

# PATENT ABSTRACTS OF JAPAN

(11)Publication number : **2000-242434**

(43)Date of publication of application : **08.09.2000**

(51)Int.Cl.

**G06F 3/06**

(21)Application number : **11-344260**

(71)Applicant : **HITACHI LTD**

(22)Date of filing : **03.12.1999**

(72)Inventor : **MATSUNAMI NAOTO**

**OEDA TAKASHI**

**YAMAMOTO AKIRA**

**AJIMATSU YASUYUKI**

**SATO MASAHIKO**

(30)Priority

Priority number : 10364079

Priority date : **22.12.1998**

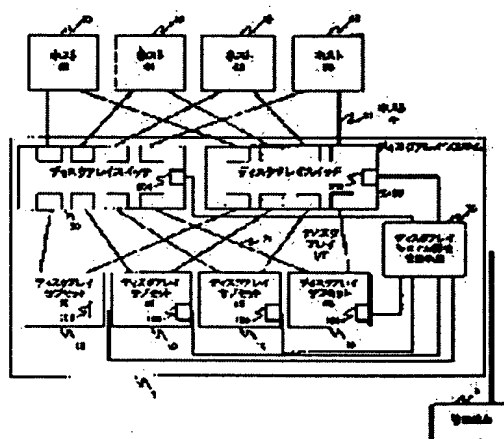
Priority country : JP

## (54) STORAGE DEVICE SYSTEM

(57)Abstract:

**PROBLEM TO BE SOLVED:** To construct a storage device system corresponding to the scale or request of a computer system so that the extension of a storage device system and improvement in reliability in the future are easily realized.

**SOLUTION:** This system 1 has a plurality of subsets 10 having a storage device for holding data and a controller for controlling the storage device and switch devices 20 arranged between the subsets 10 and a host 30. Each switch device 20 has a managing table for holding management information for managing the configuration of the storage device system 1. According to the management information, address information contained in frame information outputted by the host 30 is translated and the frame information is distributed to the subsets 10.



## LEGAL STATUS

[Date of request for examination]

24.12.2003

[Date of sending the examiner's decision of rejection]

[Kind of final disposal of application other than the examiner's decision of rejection or application converted registration]

[Date of final disposal for application]

[Patent number]

[Date of registration]

[Number of appeal against examiner's decision  
of rejection]

[Date of requesting appeal against examiner's  
decision of rejection]

[Date of extinction of right]

Copyright (C); 1998,2003 Japan Patent Office

(19) 日本国特許庁 (J P)

(12) 公開特許公報 (A)

(11) 特許出願公開番号

特開2000-242434

(P2000-242434A)

(43) 公開日 平成12年9月8日 (2000.9.8)

(51) Int.Cl.<sup>7</sup>

G 0 6 F 3/06

識別記号

3 0 1

5 4 0

F I

G 0 6 F 3/06

ターミナル\* (参考)

3 0 1 G

5 4 0

審査請求 未請求 請求項の数20 O L (全 24 頁)

(21) 出願番号 特願平11-344260

(22) 出願日 平成11年12月3日 (1999.12.3)

(31) 優先権主張番号 特願平10-364079

(32) 優先日 平成10年12月22日 (1998.12.22)

(33) 優先権主張国 日本 (J P)

(71) 出願人 000005108

株式会社日立製作所

東京都千代田区神田駿河台四丁目6番地

(72) 発明者 松並 直人

神奈川県川崎市麻生区王禅寺1099番地 株

式会社日立製作所システム開発研究所内

(72) 発明者 大枝 高

神奈川県川崎市麻生区王禅寺1099番地 株

式会社日立製作所システム開発研究所内

(74) 代理人 100075096

弁理士 作田 康夫

最終頁に続く

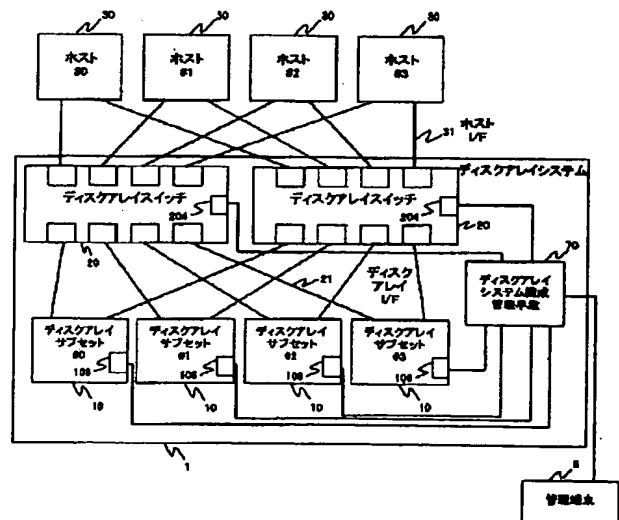
(54) 【発明の名称】 記憶装置システム

(57) 【要約】

【課題】 計算機システムの規模、要求などに応じた記憶装置システムを構築でき、将来における記憶装置システムの拡張、信頼性の向上を容易に実現できるようにする。

【解決手段】 記憶装置システム1は、データを保持する記憶装置とそれを制御する制御装置を有する複数のサブセット10とサブセット10とホスト30との間に配置されるスイッチ装置20を有する。スイッチ装置20は、記憶装置システム1の構成を管理する管理情報を保持する管理テーブルを有し、管理情報に従ってホスト30が出力するフレーム情報に含まれるアドレス情報を変換してフレーム情報をサブセット10に振り分ける。

図1



# 【特許請求の範囲】

【請求項1】データを保持する記憶媒体を有する記憶装置と該記憶装置を制御する制御装置とを有する複数の記憶装置サブシステムと、前記複数の記憶装置サブシステムに保持されるデータを使用する計算機に接続され、記憶装置システムの構成情報を格納した構成管理テーブルと、前記計算機から送られてくるフレームにตอบสนองして、該フレームを解析し、前記構成管理テーブルに保持された構成情報に基づいて前記フレームを変換するフレーム変換手段とを有する第1のインタフェースノードと、各々が前記記憶装置サブシステムのいずれか1つに接続された複数の第2のインタフェースノードと、前記第1のインタフェースノード及び前記複数の第2のインタフェースノードが接続され、前記第1のインタフェースノードと前記複数の第2のインタフェースノードとの間で前記フレームの転送を行う転送手段とを有することを特徴とする記憶装置システム。

【請求項2】前記第1のインタフェースノードは、前記フレームに前記第2のインタフェースノードのノードアドレス情報を付加して出力するパケット生成手段を有し、前記転送手段は、前記ノードアドレス情報に基づいて前記第1のインタフェースノードと前記複数の第2のインタフェースノードとの間で前記フレームの転送を行うことを特徴とする請求項1記載の記憶装置システム。

【請求項3】前記フレームは、転送元及び転送先を指定する識別子を保持するヘッダ部と、転送される実体データを保持するデータ実体部とを有し、前記変換手段は、前記構成情報に基づき前記ヘッダ部に保持された転送先の識別子を変換することを特徴とする請求項1記載の記憶装置システム。

【請求項4】前記フレームは、前記データ実体部に、前記計算機により認識されている第1の論理アドレス情報を含み、前記変換手段は、前記構成管理テーブルに保持された前記構成情報に基づいて、前記第1の論理アドレス情報を、該フレームの転送先となる記憶装置サブシステム内で管理される第2の論理アドレスに変換することを特徴とする請求項3記載の記憶装置システム。

【請求項5】前記記憶装置システムは、さらに、前記転送手段に接続し、オペレータから記憶装置システムの構成を定義する構成情報の入力を受け付け、該入力にตอบสนองして、各ノードの前記構成管理テーブルに前記構成情報を設定する管理プロセッサを有することを特徴とする請求項1記載の記憶装置システム。

【請求項6】前記構成情報は、前記計算機から前記複数の記憶装置サブシステムへのアクセスを制限する情報を含むことを特徴とする請求項5記載の記憶装置システム。

【請求項7】前記第1のインタフェースノードは、前記計算機から転送されてくるデータの書き込みを指示するライトコマンドフレームにตอบสนองして、該ライトコマンド

フレーム及びそれに続くデータフレームについてそれらの複製を生成し、前記ライトコマンドフレーム及びそれに続くデータフレームが少なくとも2つの記憶装置サブシステムに送られるよう、各々のフレームに異なるノードアドレス情報を付加して前記転送手段に転送することとを特徴とする請求項2記載の記憶装置システム。

【請求項8】前記第1のインタフェースノードは、前記計算機から転送されてくるデータのリードを指示するリードコマンドフレームにตอบสนองして、該リードコマンドフレームの複製を生成し、前記少なくとも2つの記憶装置サブシステムに前記リードコマンドフレームが送られるように、各々のリードコマンドフレームに異なるノードアドレス情報を付加して前記転送手段に転送することとを特徴とする請求項7記載の記憶装置サブシステム。

【請求項9】前記第1のインタフェースノードは、前記リードコマンドフレームにตอบสนองして前記少なくとも2つの記憶装置サブシステムから転送されてくるデータフレームを受信し、その一方を選択して前記計算機に転送することを特徴とする請求項8記載の記憶装置システム。

【請求項10】前記第1のインタフェースノードは、前記計算機から転送されてくるデータのリードを指示するリードコマンドフレームにตอบสนองして、前記少なくとも2つの記憶装置サブシステムのうち予め定められた一の記憶装置サブシステムに接続する第2のインタフェースノードのノードアドレス情報を前記リードコマンドフレームに付加して前記転送手段に転送することを特徴とする請求項7記載の記憶装置サブシステム。

【請求項11】データを保持する記憶媒体を有する記憶装置、及び該記憶装置を制御する制御装置とを有する複数の記憶装置サブシステムと、前記記憶装置に格納されたデータを利用する計算機との間に接続されるスイッチ装置であって、前記計算機に接続され、記憶装置システムの構成情報を格納した構成管理テーブルと、前記計算機から送られてくるフレームにตอบสนองして、該フレームを解析し、前記構成管理テーブルに保持された前記構成情報に基づいて前記フレームを変換する変換手段と、各々が前記記憶装置サブシステムのいずれかに接続された複数の第2のインタフェースノードと、前記第1のインタフェースノード及び前記複数の第2のインタフェースノードが接続され、前記第1のインタフェースノードと前記複数の第2のインタフェースノードとの間で前記フレームの転送を行う転送手段とを有することとを特徴とするスイッチ装置。

【請求項12】前記第1のインタフェースノードが、前記フレームに前記第2のインタフェースノードのノードアドレス情報を付加して出力するパケット生成手段を有し、前記転送手段は、前記ノードアドレス情報に基づいて前記第1のインタフェースノードと前記複数の第2のインタフェースノードとの間で前記フレームの転送を行うことを特徴とする請求項11記載のスイッチ装置。

【請求項13】前記フレームは、転送元及び転送先を指定する識別子を保持するヘッダ部と、転送される実体データを保持するデータ実体部とを有し、前記変換手段は、前記構成情報に基づき前記ヘッダ部に保持された転送先の識別子を変換することを特徴とする請求項11記載のスイッチ装置。

【請求項14】前記フレームは、前記データ実体部に、前記計算機により認識されている前記データの格納先を示す第1の論理アドレス情報を含み、前記変換手段は、前記構成管理テーブルに保持された前記構成情報に基づいて、前記第1の論理アドレス情報を、該フレームの転送先となる記憶装置サブシステム内で管理される第2の論理アドレスに変換することを特徴とする請求項13記載のスイッチ装置。

【請求項15】前記スイッチ装置は、さらに、前記転送手段に接続し、オペレータから該スイッチ装置及び前記複数の記憶装置サブシステムを含んで構成される記憶装置システムの構成を定義する構成情報の入力を受け付け、該入力にตอบสนองして、各ノードの構成管理テーブルに前記構成情報を設定する管理プロセッサを有することを特徴とする請求項11記載のスイッチ装置。

【請求項16】前記第1のインタフェースノードは、前記計算機から転送されてくるデータの書き込みを指示するライトコマンドフレームにตอบสนองして、該ライトコマンドフレーム及びそれに続くデータフレームについてそれらの複製を生成し、前記ライトコマンドフレーム及びそれに続くデータフレームが少なくとも2つの記憶装置サブシステムに送られるよう、各々のフレームに異なるノードアドレス情報を付加して前記転送手段に転送することを特徴とする請求項12記載のスイッチ装置。

【請求項17】前記第1のインタフェースノードは、前記計算機から転送されてくるデータのリードを指示するリードコマンドフレームにตอบสนองして、該リードコマンドフレームの複製を生成し、前記少なくとも2つの記憶装置サブシステムに前記リードコマンドフレームが送られるように、各々のリードコマンドフレームに異なるノードアドレス情報を付加して前記転送手段に転送することを特徴とする請求項16記載のスイッチ装置。

【請求項18】前記第1のインタフェースノードは、前記リードコマンドフレームにตอบสนองして前記少なくとも2つの記憶装置サブシステムから転送されてくるデータフレームを受信し、その一方を選択して前記計算機に転送することを特徴とする請求項17記載のスイッチ装置。

【請求項19】前記第1のインタフェースノードは、前記計算機から転送されてくるデータのリードを指示するリードコマンドフレームにตอบสนองして、前記少なくとも2つの記憶装置サブシステムのうち予め定められた一の記憶装置サブシステムに接続する第2のインタフェースノードのノードアドレス情報を前記リードコマンドフレームに付加して前記転送手段に転送することを特徴とする

請求項16記載のスイッチ装置。

【請求項20】データを保持する記憶媒体を有する記憶装置と、該記憶装置を制御する制御装置とを有する複数の記憶装置サブシステムと、前記複数の記憶装置サブシステムに保持されるデータを使用する計算機に接続された第1のインタフェースノードと、各々が前記記憶装置サブシステムのいずれか1つに接続された複数の第2のインタフェースノードと、前記第1のインタフェースノード及び前記複数の第2のインタフェースノードが接続され、前記第1のインタフェースノードと前記複数の第2のインタフェースノードとの間でフレームの転送を行う転送手段と、前記転送手段に接続し、オペレータにより入力された記憶装置システムの構成を定義する構成情報を保持する管理テーブルを備えて前記構成情報に基づいて該記憶装置システムの構成を管理する管理プロセッサとを有することを特徴とする記憶装置システム。

【発明の詳細な説明】

【0001】

【発明の属する技術分野】本発明は、複数のディスク装置を制御するディスク制御システムの実現方法に関し、特に、ディスク制御システムの高速度化、低コスト化、コストパフォーマンスの向上の方法に関する。

【0002】

【従来の技術】計算機システムに用いられる記憶装置システムとして、複数のディスク装置を制御するディスクアレイシステムがある。ディスクアレイシステムについては、例えば、“A Case for Redundant Arrays of Inexpensive Disks (RAID)”；InProc. ACM SIGMOD, June 1988（カリフォルニア大学バークレー校発行）に開示されている。ディスクアレイは、複数のディスク装置を並列に動作させることで、ディスク装置を単体で用いた記憶装置システムに比べ高速度化を実現する技術である。

【0003】複数のディスクアレイシステムを、複数のホストと相互に接続する方法として、ファイバチャネル（Fibre Channel）のFabricを使用した方法がある。この方法を適用した計算機システムの例が、日経エレクトロニクス1995.7.3（no.639）「シリアルSCSIがよいよ市場へ」P.79 図3に示されている。ここに開示される計算機システムでは、複数のホストコンピュータ（以下では単にホストと呼ぶ）と複数のディスクアレイシステムが、それぞれ、ファイバチャネルを介してファブリック装置に接続される。ファブリック装置は、ファイバチャネルのスイッチであり、ファブリック装置に接続する任意の装置間の転送路の接続を行う。ファブリック装置はファイバチャネルのパケットである「フレーム」の転送に対し透過であり、ホストとディスクアレイシステムは、互いにファブリック装置を意識することなく2点間で通信を行う。

【0004】

【発明が解決しようとする課題】従来のディスクアレイ

システムでは、大容量化のためディスク装置の台数を増やし、高性能化のため台数に見合った性能を有するコントローラを実現しようとする、コントローラの内部バスの性能限界や、転送制御を行うプロセッサの性能限界が顕在化する。このような問題に対処するために、内部バスを拡張し、プロセッサ数を増加することが行われている。しかし、このような対処の仕方は、多数のバス制御によるコントローラ構成の複雑化や、プロセッサ間の共有データの排他制御等による制御ソフトの複雑化とオーバーヘッドの増加を招く。このため、コストを非常に上昇させるとともに、性能は頭打ちになり、その結果、コストパフォーマンスが悪化する。また、このような装置は、大規模なシステムでは、そのコストに見合った性能が実現できるものの、規模がそれほど大きくないシステムには見合わない、拡張性が制限される、開発期間の増大と開発コストの上昇を招くといった課題がある。

【0005】複数のディスクアレイシステムを並べファブリック装置で相互接続することによって、システム全体としての大容量化、高性能化を行うことが可能である。しかし、この方法では、ディスクアレイシステム間に関連性は全くなく、特定のディスクアレイシステムにアクセスが集中したとしてもそれを他の装置に分散することができないので、実使用上の高性能化が実現できない。また、ホストから見た論理的なディスク装置（論理ユニットと呼ぶ）の容量は、1台のディスクアレイシステムの容量に制限されるので、論理ユニットの大容量化は実現できない。

【0006】ディスクアレイシステム全体を高信頼化しようとした際に、ホストが備えているミラーリング機能を用いて2台のディスクアレイシステムによるミラー構成を実現することができるが、ホストによるミラーリングのための制御オーバーヘッドが発生し、システム性能が制限されるという課題がある。また、多数のディスクアレイシステムがシステム内に個別に存在すると、システム管理者が管理するための負荷が増加する。このため、多数の保守人員、複数台分の保守費用が必要になる等、管理コストが増加する。さらに、複数のディスクアレイシステム、ファブリック装置は、それぞれ独立した装置であるので、各種設定は、それぞれの装置毎に異なる方法で実施する必要がある。このため、管理者のトレーニングや、操作時間の増大にともない運用コストが増大する。

【0007】本発明の目的は、これら従来技術における課題を解決し、計算機システムの規模、要求などに応じた記憶装置システムを構築でき、将来における記憶装置システムの拡張、信頼性の向上などに容易に対応することのできる記憶装置システムを実現することにある。

【0008】

【課題を解決するための手段】本発明の記憶装置システムは、データを保持する記憶媒体を有する記憶装置と、

この記憶装置を制御する制御装置とを有する複数の記憶装置サブシステム、複数の記憶装置サブシステムに保持されるデータを使用する計算機に接続された第1のインタフェースノード、各々が記憶装置サブシステムのいずれかに接続された複数の第2のインタフェースノード、及び第1のインタフェースノード及び複数の第2のインタフェースノードが接続され、第1のインタフェースノードと複数の第2のインタフェースノードとの間でフレームの転送を行う転送手段を有する。

【0009】好ましくは、第1のインタフェースノードは、記憶装置システムの構成情報を格納した構成管理テーブルと、計算機から送られてくるフレームにตอบสนองして、該フレームを解析し、構成管理テーブルに保持された構成情報に基づいてそのフレームの転送先に関する情報変換して転送手段に転送する。

【0010】また、フレームの転送に際して、第1のインタフェースノードは、そのフレームを受け取るべきノードのノードアドレス情報をフレームに付加する。転送手段はフレームに付加されたノードアドレス情報に従ってフレームを転送する。第2のインタフェースノードは、転送手段から受け取ったフレームからノードアドレス情報を除いてフレームを再形成し、目的の記憶装置サブシステムに転送する。

【0011】本発明のある態様において、記憶装置システムは、転送手段に接続する管理プロセッサを有する。管理プロセッサは、オペレータからの指示に従って、構成管理テーブルに構成情報を設定する。構成情報には、計算機からのアクセスを制限する情報が含まれる。

【0012】

【発明の実施の形態】〔第1実施形態〕図1は、本発明が適用されたディスクアレイシステムを用いたコンピュータシステムの一実施形態における構成図である。

【0013】1はディスクアレイシステム、30はディスクアレイシステムが接続されるホストコンピュータ（ホスト）である。ディスクアレイシステム1は、ディスクアレイサブセット10、ディスクアレイスイッチ20、ディスクアレイシステム全体の設定管理を行うディスクアレイシステム構成管理手段70、ディスクアレイスイッチ20とディスクアレイシステム構成管理手段70との間、およびディスクアレイサブセット10とディスクアレイシステム構成管理手段70との間の通信インタフェース（通信I/F）80を有する。ホスト30とディスクアレイシステム1とは、ホストインタフェース（ホストI/F）31で接続されており、ホストI/F31はディスクアレイシステム1のディスクアレイスイッチ20に接続する。ディスクアレイシステム1の内部において、ディスクアレイスイッチ20とディスクアレイサブセット10は、ディスクアレイインタフェース（ディスクアレイI/F21）で接続される。

【0014】ホスト30、ディスクアレイサブセット1

0は、図では、各々4台示されているが、この台数に関しては制限はなく任意である。ホスト30とディスクアレイサブセット10の台数が異なっても構わない。また、ディスクアレイスイッチ20は、本実施形態では図示の通り二重化されている。各ホスト30および各ディスクアレイサブセット10は、それぞれ別々のホストI/F31、ディスクアレイI/F21で二重化されたディスクアレイスイッチ20の双方に接続されている。これは、一方のディスクアレイスイッチ20、ホストI/F31、あるいはディスクアレイI/F21が故障しても他方を使用することでホスト30からディスクアレイシステム1へのアクセスを可能とし、高い可用性を実現するためである。しかし、このような二重化は必ずしも必須ではなく、システムに要求される信頼性レベルに応じて選択可能である。

【0015】図2は、ディスクアレイサブセット10の一構成例を示す構成図である。101は上位システム（ホスト10）からのコマンドを解釈してキャッシュヒットミス判定を実施し、上位システムとキャッシュ間のデータ転送を制御する上位アダプタ、102はディスクデータアクセス高速化のためのキャッシュ、および、マルチプロセッサ間の共有データを格納する共有メモリ（以下キャッシュ・共有メモリと呼ぶ）、104はディスクアレイサブセット10内に格納される複数のディスクユニットである。103はディスクユニット104を制御し、ディスクユニット104とキャッシュ間のデータ転送を制御する下位アダプタである。106はディスクアレイサブセット構成管理手段であり、ディスクアレイシステム1全体を管理するディスクアレイシステム構成管理手段70と通信I/F80を介して通信し、構成パラメータの設定や、障害情報の通報等の管理を行う。

【0016】上位アダプタ101、キャッシュ・共有メモリ102、下位アダプタ103はそれぞれ二重化されている。この理由は上記ディスクアレイスイッチ20の二重化と同様、高可用性を実現するためであり必須ではない。また、各ディスクユニット104は、二重化された下位アダプタ103のいずれからも制御可能である。本実施形態では、低コスト化の観点から同一のメモリ手段をキャッシュと共有メモリに共用しているが、これらは勿論分離することも可能である。

【0017】上位アダプタ101は、上位アダプタ101の制御を実行する上位MPU1010、上位システム、すなわちディスクアレイスイッチ20との接続I/FであるディスクアレイI/F21を制御するディスクアレイI/Fコントローラ1011、キャッシュ・共有メモリ102と上位MPU1010とディスクアレイI/Fコントローラ1011との間の通信、データ転送を行う上位バス1012を含む。

【0018】図では各上位アダプタ101毎に1台のディスクアレイI/Fコントローラ1011が示されてい

るが、1つの上位アダプタに対し、複数のディスクアレイI/Fコントローラ1011を設けてもよい。

【0019】下位アダプタ103は、下位アダプタ103の制御を実行する下位MPU1030、ディスク104とのインタフェースであるディスクI/Fを制御するディスクI/Fコントローラ1031、キャッシュ・共有メモリ102と下位MPU1030とディスクI/Fコントローラ1031との間の通信、データ転送を行う下位バス1032を含む。

【0020】図では各下位アダプタ103毎に4台のディスクI/Fコントローラ1031が示されているが、その数は任意であり、ディスクアレイの構成や、接続するディスク台数に応じて変更可能である。

【0021】図3は、ディスクアレイスイッチ20の一構成例を示す構成図である。200はディスクアレイスイッチ全体の制御および管理を行うプロセッサである管理プロセッサ（MP）、201は $n \times n$ の相互スイッチ経路を構成するクロスバススイッチ、202はディスクアレイI/F21毎に設けられるディスクアレイI/Fノード、203はホストI/F31毎に設けられるホストI/Fノード、204はディスクアレイシステム構成管理手段70との間の通信を行う通信コントローラである。2020はディスクアレイI/Fノード202とクロスバススイッチ201を接続するバス、2030はホストI/Fノード203とクロスバススイッチ201を接続するバス、2040は他のディスクアレイスイッチ20と接続し、クラスタを構成するためのクラスタ間I/F、2050はMP200とクロスバススイッチ201を接続するためのバスである。

【0022】図4はクロスバススイッチ201の構造を示す構成図である。2010はクロスバススイッチ201に接続するバス2020、2030、2050、およびクラスタ間I/F2040を接続するポートであるスイッチングポート（SWP）である。SWP2010はすべて同一の構造を有し、あるSWPから他のSWPへの転送経路のスイッチング制御を行う。図では1つのSWPについてのみ転送経路を示しているが、すべてのSWP間で同様の転送経路が存在する。

【0023】図5は、ホストI/Fノード203の一構成例を示す構成図である。本実施形態では、具体的に説明をするためにホストI/F31とディスクアレイI/F21の両方にファイバチャネルを使用するものと仮定する。もちろんホストI/F31とディスクアレイI/F21として、ファイバチャネル以外のインタフェースを適用することも可能である。ホストI/Fノード203とディスクアレイI/Fノード202の両方に同一のインタフェースを使用することで、両者を同一構造にできる。本実施形態においては、ディスクアレイI/Fノード202も図に示すホストI/Fノード203と同様に構成される。以下では、ホストI/Fノード203を

例に説明を行う。

【0024】2021は受信したファイバチャネルフレーム（以下単にフレームと呼ぶ）をどのノードに転送するかを検索する検索プロセッサ（SP）、2022はホスト30（ディスクアレイI/Fノード202の場合）、ディスクアレイサブセット10）との間でフレームを送受信するインタフェースコントローラ（IC）、2022はIC2023が受信したフレームに対しSP2021が検索した結果に基づいて変換を施すスイッチングコントローラ（SC）、2024はSC2021が変換したフレームを他のノードに転送するためにクロスバスイッチ201を通過できる形式にパケット化するパケット生成部（SPG）、2025は受信したフレームを一時的に格納するフレームバッファ（FB）、2026は一つのホストからのディスクアレイアクセス要求コマンド（以下単にコマンドと呼ぶ）に対応した複数のフレーム列であるエクステンジ（Exchange）を識別するためのエクステンジ番号を管理するエクステンジテーブル（ET）、2027は複数のディスクアレイサブセット10の構成情報を格納するディスクアレイ構成管理テーブル（DCT）である。

【0025】ディスクアレイスイッチ20の各構成部は、すべてハードウェアロジックで構成されることが性能上望ましい。しかし、求められる性能を満足できるならば、汎用プロセッサを用いたプログラム制御によりSP2021やSC2022の機能を実現することも可能である。

【0026】各ディスクアレイサブセット10は、各々が有するディスクユニット104を1または複数の論理的なディスクユニットとして管理している。この論理的なディスクユニットを論理ユニット（LU）と呼ぶ。LUは、物理的なディスクユニット104と1対1で対応する必要はなく、1台のディスクユニット104に複数のLUが構成され、あるいは、複数のディスクユニット104で1つのLUが構成されても構わない。

【0027】ディスクアレイサブセット10の外部から見た場合、1つのLUは、1台のディスク装置として認識される。本実施形態では、ディスクアレイスイッチ20によりさらに論理的なLUが構成され、ホスト30は、このLUに対してアクセスするように動作する。本明細書では、1つのLUでホスト30から認識される1つのLUが構成される場合、ホスト30により認識されるLUを独立LU（ILU）、複数のLUでホスト30から認識される1つのLUが構成される場合、ホスト30により認識されるLUを統合LU（CLU）と呼ぶ。

【0028】図12に、4つのディスクアレイサブセットのLUで1つの統合LUが構成される場合における各階層間でのアドレス空間の対応関係を示す。図において、1000は、一例として、ホスト“#2”からみたディスクアレイシステム1の1つの統合LUにおけるア

ドレス空間、1100は、ディスクアレイサブセット10のLUのアドレス空間、1200はディスクユニット104（ここでは、ディスクアレイサブセット“#0”についてのみ図示されている）のアドレス空間を示している。

【0029】各ディスクアレイサブセット10のLUは、ここでは、4台のディスクユニット104によりRAID5（Redundant Arrays of Inexpensive Disks Level 5）型ディスクアレイとして構成されるものとする。各ディスクアレイサブセット10は、それぞれn0、n1、n2、n3の容量を有するLUを持つ。ディスクアレイスイッチ20は、これら4つのLUの持つアドレス空間を（n0+n1+n2+n3）の容量を有するアドレス空間に統合し、ホスト30から認識される統合LUを実現する。

【0030】本実施形態では、例えば、ホスト#2が領域A1001をアクセスする場合、領域A1001を指定したアクセス要求は、ディスクアレイスイッチ20によりディスクアレイサブセット#0のLUの領域A'1101をアクセスするための要求に変換されてディスクアレイサブセット#0に転送される。ディスクアレイサブセット#0は、領域A'1101をさらに、ディスクユニット104上の領域A''1201にマッピングしてアクセスを行う。アドレス空間1000とアドレス空間1100との間のマッピングは、ディスクアレイスイッチ20が有するDCT207に保持された構成情報に基づき行われる。この処理の詳細については後述する。なお、ディスクアレイサブセット内におけるマッピングについては、既によく知られた技術であり、本明細書では詳細な説明については省略する。

【0031】本実施形態において、DCT207は、システム構成テーブルとサブセット構成テーブルを含む。図6は、システム構成テーブルの構成を、図7は、サブセット構成テーブルの構成を示す。

【0032】図7に示すように、システム構成テーブル20270は、ホストLUの構成を示す情報を保持するホストLU構成テーブル20271、及びディスクアレイスイッチ20のディスクアレイI/Fノード202とディスクアレイサブセット10との接続関係を示すディスクアレイI/Fノード構成テーブル20272を有する。

【0033】ホストLU構成テーブル20271は、ホスト30からみたLUごとに、そのLUを識別する番号であるHost-LU No.、LUの属性を示すLU Type、CLU Class、及びCLU Stripe Size、ホストLUの状態を示す情報であるCondition、ホストLUを構成するディスクアレイサブセット10のLUに関する情報であるLU情報（LU Info.）を有する。

【0034】LU Typeは、このホストLUがCLUであるか、ILUであるかといったLUの種類を示す情報で



ある。CLU Classは、LU TypeによりこのホストLUがCLUであることが示される場合に、そのクラスが“Joined”、“mirrored”、及び“Striped”のいずれであるかを示す情報である。“Joined”は、図11により説明したように、いくつかのLUを連結して1つの大きな記憶空間を持つCLUが構成されていることを示す。“Mirrored”は、第6実施形態として後述するように、2つのLUにより二重化されたLUであることを示す。“Striped”は、第7実施形態として後述するように、複数のLUで構成され、データがこれら複数のLUに分散して格納されたLUであることを示す。CLU Stripe Sizeは、CLU Classにより「Striped」であることが示される場合に、ストライピングサイズ（データの分散の単位となるブロックのサイズ）を示す。

【0035】Conditionにより示される状態には、“Normal”、“Warning”、“Fault”、及び“Not Defined”の4種類がある。“Normal”はこのホストLUが正常な状態であることを示す。“Warning”は、このホストLUを構成するLUに対応するいずれかのディスクユニットに障害が発生している等の理由により縮退運転が行われていることを示す。“Fault”は、ディスクアレイサブセット10の故障などによりこのホストLUを運転することができないことを示す。“Not Defined”は、対応するHost-LU No.のホストLUが定義されていないことを示す。

【0036】LU Infoは、このホストLUを構成するLUについて、そのLUが属するディスクアレイサブセット10を特定する情報、ディスクアレイサブセット内でのLUN、及びそのサイズを示す情報を含む。ホストLUがILUの場合には、唯一のLUに関する情報が登録される。ホストLUがCLUの場合には、それを構成する全てのLUについて、それぞれのLUに関する情報が登録される。例えば、図において、Host-LU No.が“0”であるHost-LUは、ディスクアレイサブセット“#0”のLUN“0”、ディスクアレイサブセット“#1”のLUN“0”、ディスクアレイサブセット“#2”のLUN“0”、ディスクアレイサブセット“#3”のLUN“0”の4つのLUから構成されるCLUであり、そのCLUクラスが“Joined”であるCLUであることが分かる。

【0037】ディスクアレイI/Fノード構成テーブル20272は、ディスクアレイI/F21が接続するディスクアレイサブセット10のポートごとに、どのディスクアレイスイッチ20のディスクアレイI/Fノード202が接続されるかを示す情報を保持する。

【0038】具体的には、ディスクアレイサブセット10を特定するSubset No.、ポートを特定するSubset Port No.、そのポートに接続するディスクアレイスイッチ20を特定するSwitch No.、及びそのディスクアレイスイッチ20のディスクアレイI/Fノード202を特定

するI/F Node No.を有する。ディスクアレイサブセット10が複数のポートを備えている場合には、そのポート毎に情報が設定される。

【0039】サブセット構成テーブルは、図7に示すように、各ディスクアレイサブセット10に対応する複数のテーブル202720～202723を有する。各テーブルは、ディスクアレイサブセット10内で構築されたRAIDグループの構成を示す情報を保持するRAIDグループ構成テーブル202730と、ディスクアレイサブセット10内に構築されたLUの構成を示す情報を保持するLU構成テーブル202740を含む。

【0040】RAIDグループ構成テーブル202730は、RAIDグループに付加された番号を示すGroup No.、そのRAIDグループのレベルを示すLevel、そのRAIDグループを構成するディスクの数を示す情報であるDisks、そのRAIDグループがRAIDレベル0、5等のストライピングされた構成の場合、そのストライピングサイズを示すStripe Sizeを情報として含む。例えば、図に示されるテーブルにおいて、RAIDグループ“0”は、4台のディスクユニットにより構成されたRAIDグループであり、RAIDレベルが5、ストライピングサイズがS0である。

【0041】LU構成テーブル202740は、LUに付加された番号（LUN）を示すLU No.、このLUがどのRAIDグループに構成されているのかを示すRAID Group、LUの状態を示すCondition、このLUのサイズ（容量）を示すSize、このLUがディスクアレイサブセット10のどのポートからアクセス可能なかを示すPort、及びその代替となるポートを示すAlt. Portを情報として含む。Conditionで示される状態は、ホストLUについてのConditionと同様、“Normal”、“Warning”、“Fault”、“Not Defined”の4種類がある。Alt. Portに設定された情報により特定されるポートは、Portに設定された情報で特定されるポートに障害が発生したときに用いられるが、単に複数のポートから同一のLUをアクセスするために用いることもできる。

【0042】図8は、ファイバチャネルにおけるフレームの構成図である。ファイバチャネルのフレーム400は、フレームの先頭を示すSOF (Start Of Frame) 400、フレームヘッダ401、転送の実態データを格納する部位であるフレームペイロード402、32ビットのエラー検出コードであるCRC (Cyclic Redundancy Check) 403、フレームの最後尾を示すEOF (End Of Frame) 404を含む。フレームヘッダ401は、図9に示すような構造になっており、フレーム転送元のID (S\_ID)、フレーム転送先のID (D\_ID)、エクステンジの起動元、応答先が指定するそれぞれのエクステンジID (OX\_ID、RX\_ID)、エクステンジ中のフレームグループを指定するシーケンスのID (SEQ\_ID) 等が格納されている。

【0043】本実施形態では、ホスト30により発行されるフレームには、S\_IDとしてホスト30に割り当てられたIDが、また、D\_IDとしてディスクアレイスイッチ20のポートに割り当てられたIDが使用される。一つのホストコマンドに対し、1ペアのエキスチェンジID (OX\_ID、RX\_ID) が割り当てられる。複数のデータフレームを同一のエキスチェンジに対し発行する必要があるときは、その全データフレームに対して同一のSEQ\_IDが割り当てられ、おのおのはシーケンスカウント (SEQ\_CNT) で識別される。フレームペイロード402の最大長は2110バイトであり、フレーム種毎に格納される内容が異なる。例えば、後述するFCP\_CMDフレームの場合、図10に示すように、SCSIのLogical Unit Number (LUN)、Command Description Block (CDB) 等が格納される。CDBは、ディスク (ディスクアレイ) アクセスに必要なコマンドバイト、転送開始論理アドレス (LBA)、転送長 (LEN) を含む。

【0044】以下、本実施形態のディスクアレイシステムの動作を説明する。

【0045】ディスクアレイシステムを使用するのに先立ち、ディスクアレイスイッチ20に対して、ディスクアレイサブセット10の構成情報を設定する必要がある。システム管理者は、管理端末5からディスクアレイシステム構成手段70を介して、すべてのディスクアレイサブセット10およびディスクアレイスイッチ20の構成設定情報を獲得する。管理者は、管理端末5から所望のシステム構成になるよう論理ユニットの構成設定、RAIDレベルの設定、障害発生時の交代パスの設定等、各種設定に必要な設定情報を入力する。ディスクアレイシステム構成管理手段70は、その設定情報を受け、各ディスクアレイサブセット10およびディスクアレイスイッチ20に設定情報を転送する。なお、管理端末5における設定情報の入力については第5実施形態にて別途説明する。

【0046】ディスクアレイスイッチ20では、通信コントローラ204が設定情報を獲得し、MP200により各ディスクアレイサブセット10のアドレス空間情報等の構成情報が設定される。MP200は、クロスバススイッチ201経由で各ホストI/Fノード203およびディスクアレイI/Fノード202に、ディスクアレイサブセット10の構成情報を配信する。

【0047】各ノード203、および202はこの情報を受信すると、SP2021により構成情報をDCT2027に格納する。ディスクアレイサブセット10では、ディスクアレイサブセット構成管理手段106が、設定情報を獲得し、共有メモリ102に格納する。各上位MPU1010および下位MPU1030は、共有メモリ102上の設定情報を参照し、各々の構成管理を実施する。

【0048】以下では、ホスト“#2”がディスクアレイ

システム1に対し、リードコマンドを発行した場合の動作を説明する。図11に、ホストからのリード動作時にファイバチャネルを通して転送されるフレームのシーケンスを示す模式図を、図13にこのときのディスクアレイスイッチのホストI/Fノード203における動作のフローチャートを示す。

【0049】なお、以下の説明では、ホスト“#2”が、図12における記憶領域A1001をアクセスすることを仮定する。記憶領域A1001に対応する実際の記憶領域Aは、ディスクアレイサブセット“#0”のLUN=0のLUを構成するディスクユニット#2のアドレス空間内に存在するものとする。また、アドレス空間1000を構成するLUを定義しているホストLU構成テーブル20271のLU Typeには「CLU」が、CLU Classには「Joined」が設定されているものとする。

【0050】データのリード時、ホスト30は、リードコマンドを格納したコマンドフレーム「FCP\_CMD」をディスクアレイスイッチ20に発行する (図11矢印(a))。ディスクアレイスイッチ20のホストI/Fノード“#2”は、IC2023によりホストI/F31経由でコマンドフレーム「FCP\_CMD」を受信する (ステップ20001)。IC2023は、SC2022にコマンドフレームを転送する。SC2022は、受け取ったコマンドフレームを一旦FB2025に格納する。この際、SC2022は、コマンドフレームのCRCを計算し、受信情報が正しいことを検査する。CRCの検査に誤りがあれば、SC2022は、その旨をIC2023に通知する。IC2023は、誤りの通知をSC2022から受けると、ホストI/F31を介してホスト30にCRCエラーを報告する。 (ステップ20002)。

【0051】CRCが正しい場合、SC2022は、FB2025に保持したフレームをリードし、それがコマンドフレームであることを認識してフレームヘッダ401を解析する (ステップ20003)。そして、SC2022は、SP2021に指示し、S\_ID、D\_ID、OX\_ID等のエキスチェンジ情報をET2026に登録する (ステップ20004)。

【0052】次に、SC2022は、フレームペイロード402を解析し、ホスト30により指定されたLUNおよびCDBを取得する (ステップ20005)。SP2021は、SC2022の指示により、DCT2027を検索し、ディスクアレイサブセット10の構成情報を得る。具体的には、SP2021は、ホストLU構成テーブル20271を検索し、受信したフレームペイロード402に格納されたLUNと一致するHost-LU No.を有する情報を見つける。SP2021は、LU Type、CLU Classに設定された情報からホストLUの構成を認識し、LU Info.に保持されている情報に基づきアクセスすべきディスクサブセット10とその中のLUのLUN、及びこのLU内でのLBAを判別する。次に、SP2021は、

サブセット構成テーブル202720のLU構成テーブル202740を参照し、目的のディスクアレイサブセット10の接続ポートを確認し、ディスクアレイI/Fノード構成テーブル20272からそのポートに接続するディスクアレイI/Fノード202のノードNo.を得る。SP2021は、このようにして得たディスクアレイサブセット10を識別する番号、LUN、LBA等の変換情報をSC2022に報告する。(ステップ20006)。

【0053】次に、SC2022は、獲得した変換情報を使用しフレームペイロード402のLUNとCDBのなかのLBAを変換する。また、フレームヘッダ401のD\_IDを対応するディスクアレイサブセット10のホストI/Fコントローラ1011のD\_IDに変換する。なお、この時点ではS\_IDは書き換えない(ステップ20007)。

【0054】SC2022は、変換後のコマンドフレームと、対象ディスクアレイサブセット10に接続するディスクアレイI/Fノード番号を、SPG2024に転送する。SPG2024は、受け取った変換後のコマンドフレームに対し、図14に示すような簡単な拡張ヘッダ601を付加したパケットを生成する。このパケットをスイッチングパケット(S Packet)60と呼ぶ。S Packet60の拡張ヘッダ601には、転送元(自ノード)番号、転送先ノード番号、及び転送長が付加含まれる。SPG2024は、生成したS Packet60をクロスバスイッチ201に送信する(ステップ20008)。

【0055】クロスバスイッチ201は、ホストI/Fノード“#2”と接続するSWP2010によりS Packet60を受信する。SWP2010は、S Packet60の拡張ヘッダ601を参照し、転送先のノードが接続するSWPへのスイッチ制御を行って経路を確立し、S Packet60を転送先のディスクアレイI/Fノード202(ここでは、ディスクアレイI/Fノード“#0”)に転送する。SWP2010は、経路の確立をS Packet60の受信の度に実施し、S Packet60の転送が終了したら、その経路を解放する。ディスクアレイI/Fノード“#0”では、SPG2024がS Packet60を受信し、拡張ヘッダ601を外してコマンドフレームの部分をSC2022に渡す。

【0056】SC2022は、受け取ったコマンドフレームのフレームヘッダのS\_IDに自分のIDを書き込む。次にSC2022は、SP2021に対し、コマンドフレームのS\_ID、D\_ID、OX\_ID等のエクステンション情報、及びフレーム転送元ホストI/Fノード番号をET2026に登録するよう指示し、IC2023にコマンドフレームを転送する。IC2023は、フレームヘッダ401の情報に従い、接続するディスクアレイサブセット10(ここでは、ディスクアレイサブセット“#0”)にコマンドフレームを転送する(図11矢印(b))。

【0057】ディスクアレイサブセット“#0”は、変

換後のコマンドフレーム「FCP\_CMD」をディスクアレイI/Fコントローラ1011で受信する。上位MPU1010は、コマンドフレームのフレームペイロード402に格納されたLUNとCDBを取得し、指定された論理ユニットのLBAからLEN長のデータをリードするコマンドであると認識する。

【0058】上位MPU1010は、共有メモリ102に格納されたキャッシュ管理情報を参照し、キャッシュヒットミス/ヒット判定を行う。ヒットすればキャッシュ102からデータ転送を実施する。ミスの場合、ディスクユニットからデータをリードする必要があるため、RAID5の構成に基づくアドレス変換を実施し、キャッシュ空間を確保する。そして、ディスクユニット2からのリード処理に必要な処理情報を生成し、下位MPU1030に処理を引き継ぐべく、共有メモリ102に処理情報を格納する。

【0059】下位MPU1030は、共有メモリ102に処理情報が格納されたことを契機に処理を開始する。下位MPU1030は、適切なディスクI/Fコントローラ1031を特定し、ディスクユニット2へのリードコマンドを生成して、ディスクI/Fコントローラ1031にコマンドを発行する。ディスクI/Fコントローラ1031は、ディスクユニット2からリードしたデータをキャッシュ102の指定されたアドレスに格納して下位MPU1030に終了報告を通知する。下位MPU1030は、処理が正しく終了したことを上位MPU1010に通知すべく共有メモリ102に処理終了情報を格納する。

【0060】上位MPU1010は、共有メモリ102に処理終了情報が格納されたことを契機に処理を再開し、ディスクアレイI/Fコントローラ1011にリードデータ準備完了を通知する。ディスクアレイI/Fコントローラ1011は、ディスクアレイスイッチ20の当該ディスクアレイI/Fノード“#0”に対し、ファイバチャネルにおけるデータ転送準備完了フレームである「FCP\_XFER\_RDY」を発行する(図11矢印(c))。

【0061】ディスクアレイI/Fノード“#0”では、データ転送準備完了フレーム「FCP\_XFER\_RDY」を受信すると、SC2022が、ディスクアレイサブセット20から受信した応答先エクステンションID(RX\_ID)を獲得し、S\_ID、D\_ID、OX\_IDを指定して、SP2021に指示しET2026の当該エクステンション情報にRX\_IDに登録する。SC2022は、データ転送準備完了フレームの転送先(コマンドフレームの転送元)のホストI/Fノード番号を獲得する。SC2022は、このフレームのS\_IDを無効化し、SPG2024に転送する。SPG2024は、先に述べたようにしてS Packetを生成し、クロスバスイッチ201経由で対象ホストI/Fノード“#2”に転送する。

【0062】ホストI/Fノード“#2”では、SPG

2024がデータ転送準備完了フレームのS Packetを受信すると、S Packetの拡張ヘッダを外し「FCP\_XFER\_RDY」を再生してSC2022に渡す(ステップ20011)。SC2022は、SP2021に指示しET2026をサーチして該当するエクスチェンジを特定する(ステップ20012)。

【0063】次に、SC2022は、フレームが「FCP\_XFER\_RDY」であるかどうか調べ(ステップ20013)、「FCP\_XFER\_EDY」であれば、ET2026の応答先エクスチェンジID(RX\_ID)の更新をSP2021に指示する。応答先エクスチェンジIDとしては、このフレームに付加されていた値が使用される(ステップ20014)。そして、SC2022は、フレームヘッダ401のS\_ID、D\_IDをホストI/Fノード203のIDとホスト30のIDを用いた適切な値に変換する(ステップ20015)。これらの処理によりフレームヘッダ401は、ホスト「#2」に対するフレームに変換される。IC2023は、ホスト「#2」に対し、このデータ転送準備完了フレーム「FCP\_XFER\_RDY」を発行する(図11の矢印(d):ステップ20016)。

【0064】ディスクアレイサブセット「#0」のディスクアレイI/Fコントローラ1011は、データ転送を行うため、データフレーム「FCP\_DATA」を生成し、ディスクアレイスイッチ20に転送する(図11矢印(e))。フレームペイロードの転送長には制限があるため、1フレームで転送できる最大のデータ長は2KBである。データ長がこれを越える場合は、必要数だけデータフレームを生成し発行する。すべてのデータフレームには同一のSEQ\_IDが割り当てられる。データフレームの発行は、同一のSEQ\_IDに対し複数のフレームが生成されることを除き(すなわちSEQ\_CNTが変化する)、データ転送準備完了フレームの場合と同様である。

【0065】ディスクアレイスイッチ20は、データ転送準備完了フレームの処理と同様に、データフレーム「FCP\_DATA」のフレームヘッダ401の変換を実施する。ただし、データフレームの転送の場合、RX\_IDが既に確立されているので、データ転送準備完了フレームの処理におけるステップ20014の処理はスキップされる。フレームヘッダ401の変換後、ディスクアレイスイッチ20は、ホスト「#2」にデータフレームを転送する(図11矢印(f))。

【0066】次に、ディスクアレイサブセット「#0」のディスクアレイI/Fコントローラ1011は、終了ステータス転送を行うため、ステータスフレーム「FCP\_RSP」を生成し、ディスクアレイスイッチ20に対し発行する(図11矢印(g))。ディスクアレイスイッチ20では、データ転送準備完了フレームの処理と同様に、SPG2024がS Packetから拡張ヘッダを外し「FCP\_RSP」ステータスフレームを再現し(ステップ20021)、SP2021によりET2026を検索しエクス

チェンジ情報を獲得する(ステップ20022)。SC2022は、その情報に基づきフレームを変換する(ステップ20023)。変換されたフレームは、IC2023によりホスト「#2」に転送される(図11矢印(h):ステップ20024)。最後にSP2021は、ET2026からエクスチェンジ情報を削除する(ステップ20025)。

【0067】以上のようにしてディスクアレイからのリード処理が行われる。ディスクアレイシステム1に対するライト処理についてもデータフレームの転送方向が逆転するのみで、上述したリード処理と同様の処理が行われる。

【0068】図3に示したように、ディスクアレイスイッチ20は、クロスバスイッチ201にクラスタ間I/F2040を備えている。図1に示したシステム構成では、クラスタ間I/F2040は使用されていない。本実施形態のディスクアレイスイッチ20は、クラスタ間I/F2040を利用して図15に示すように、他のディスクアレイスイッチと相互に接続されることができ

る。

【0069】本実施形態におけるディスクアレイスイッチ20単独では、ホスト30とディスクアレイサブセット10を合計8台までしか接続できないが、クラスタ間I/F2040を利用して複数のディスクアレイスイッチを相互接続し、接続できるホスト10とディスクアレイの数を増やすことができる。例えば、図15に示すシステムでは、4台のディスクアレイスイッチ20を使ってホスト30とディスクアレイサブセット10を合計32台まで接続でき、これら間で相互にデータ転送が可能になる。

【0070】このように、本実施形態では、ディスク容量や性能の必要性に合わせて、ディスクアレイサブセットやホストの接続台数を増加していくことができる。また、必要な転送帯域分のホストI/Fを用いてホスト-ディスクアレイシステム間を接続することができるので、容量、性能、接続台数の拡張性を大幅に向上させることができる。

【0071】以上説明した実施形態によれば、1台のディスクアレイサブセットの性能が、内部のMPUや内部バスで制限されたとしても、複数のディスクアレイサブセットを用いて、ディスクアレイスイッチによりホストとディスクアレイサブセット間を相互接続することができる。これにより、ディスクアレイシステムトータルとして高い性能を実現することができる。ディスクアレイサブセットの性能が比較的低いものであっても、複数のディスクアレイサブセットを用いることで高性能化を実現できる。したがって、低コストのディスクアレイサブセットをコンピュータシステムの規模に合わせて必要な台数だけ接続することができ、規模に応じた適切なコストでディスクアレイシステムを構築することが可能とな

る。

【0072】また、ディスク容量の増大や性能の向上が必要になったときは、ディスクアレイサブセットを必要だけ追加すればよい。さらに、複数のディスクアレイスイッチを用いて任意の数のホスト及びディスクアレイサブセットを接続できるので、容量、性能、接続台数のいずれをも大幅に向上させることができ、高い拡張性を有するシステムが実現できる。

【0073】さらにまた、本実施形態によれば、ディスクアレイサブセットとして、従来のディスクアレイシステムそのものの縮小機を用いることができるので、既に開発した大規模な制御ソフトウェア資産をそのまま利用でき、開発コストの低減と開発期間の短縮を実現することができる。

【0074】〔第2実施形態〕図16は、本発明の第2の実施形態におけるコンピュータシステムの構成図である。本実施形態は、ディスクアレイスイッチのホストI/Fノードにおいて、フレームヘッダ401のみを変換し、フレームペイロード402は操作しない点、及び、ディスクアレイスイッチ、ホストI/F、ディスクアレイI/Fが二重化されていない点で第1実施形態と構成上相違する。したがって、各部の構成は、第1実施形態と大きく変わるところがなく、その詳細については説明を省略する。

【0075】図16において、各ディスクアレイサブセット10は、複数の論理ユニット(LU)110で構成されている。各LU110は、独立LUとして構成される。一般に、各ディスクアレイサブセット10内のLU110に割り当てられるLUNは、0から始まる連続番号である。このため、ホスト30に対して、ディスクアレイシステム1内のすべてのLU110のLUNを連続的に見せる場合には、第1実施形態と同様に、フレームペイロード402のLUNフィールドを変換する必要がある。本実施形態では、各ディスクアレイサブセット10のLUNをそのままホスト30に見せることで、フレームペイロード402の変換を不要とし、ディスクアレイスイッチの制御を簡単なものとしている。

【0076】本実施形態のディスクアレイスイッチ20は、ホストI/Fノード203ごとに特定のディスクアレイサブセット10をアクセスできるものと仮定する。この場合、一つのホストI/F31を使うと、1台のディスクアレイサブセット10にあるLU110のみがアクセス可能である。1台のホストから複数のディスクアレイサブセット10のLU110をアクセスしたい場合には、そのホストを複数のホストI/Fノード203に接続する。また、複数のホスト30から1台のディスクアレイサブセット10のLU110をアクセスできるようにする場合は、同一のホストI/Fノード203にループポートや、ファブリックポート等を用い、複数のホスト30を接続する。このように構成すると、

1台のホスト30から1つのLU110をアクセスする際に、ホストI/Fノード203のD\_ID毎にディスクアレイサブセット10が確定することになるため、各LUのLUNをそのままホスト30に見せることが可能である。

【0077】本実施形態では、上述した理由により、ホスト30に、各ディスクアレイサブセット10内のLU110のLUNをそのままホスト30に見せているため、ディスクアレイスイッチ20におけるLUNの変換は不要となる。このため、ディスクアレイスイッチ20は、ホスト30からフレームを受信すると、フレームヘッダ401のみを第1実施例と同様にして変換し、フレームペイロード402は変換せずにディスクアレイサブセット10に転送する。本実施形態における各部の動作は、フレームペイロード402の変換が行われないことを除くと第1実施形態と同様であるので、ここでは詳細な説明を省略する。本実施形態によれば、ディスクアレイスイッチ20の開発を容易にできる。

【0078】〔第3実施形態〕第2実施形態では、ディスクアレイスイッチのホストI/Fノードにおいて、フレームヘッダのみを変換しているが、以下に説明する第3実施形態ではフレームヘッダも含め、フレームの変換を行わない形態について説明する。本実施形態のコンピュータシステムは、図1に示す第1実施形態におけるコンピュータシステムと同様に構成される。

【0079】第1、および第2実施形態では、ホスト30に対し、ディスクアレイサブセット10の台数や、LU110の構成等、ディスクアレイシステム1の内部構成を隠蔽している。このため、ホスト30からはディスクアレイシステム1が全体で1つの記憶装置として見える。これに対し、本実施形態では、ディスクアレイサブセット10をそのままホスト30に公開し、ホスト30がフレームヘッダのD\_IDとして直接ディスクアレイサブセットのポートのIDを使えるようにする。これにより、ディスクアレイスイッチは、フレームヘッダの情報に従ってフレームの転送を制御するだけで済み、従来技術におけるファイバチャネルのファブリック装置と同等のスイッチ装置をディスクアレイスイッチ20に替えて利用することができる。

【0080】ディスクアレイシステム構成管理手段70は、ディスクアレイサブセット10の通信コントローラ106、及びディスクアレイスイッチ20の通信手段204と通信して各ディスクアレイサブセット10及びディスクアレイスイッチ20の構成情報を獲得し、あるいは、設定する。

【0081】ディスクアレイスイッチ20は、基本的には図3に示す第1実施形態におけるディスクアレイスイッチと同様の構成を有する。しかし、本実施形態では、ホスト30が発行するフレームのフレームヘッダの情報をそのまま使ってフレームの転送を制御するため、第1

実施形態、あるいは第2実施形態でディスクアレイスイッチ20のホストI/Fノード203、ディスクアレイI/Fノード202が有するDCT2027や、SC2022、SPG2024等により実現されるフレームヘッダ等の変換の機能は不要となる。ディスクアレイスイッチ20が有するクロスバスイッチ201は、フレームヘッダの情報に従ってホストI/Fノード203、及びディスクアレイI/Fノード202の間でファイバチャネルのフレームの転送を行う。

【0082】本実施形態では、ディスクアレイシステムの構成をディスクアレイシステム構成管理手段70で一括して管理するために、ディスクアレイ管理用テーブル（以下、このテーブルもDCTと呼ぶ）をディスクアレイシステム構成管理手段70に備える。ディスクアレイシステム構成管理手段70が備えるDCTは、図6、7に示す、システム構成テーブル20270とサブセット構成テーブル202720～202723の2つのテーブル群を含む。なお、本実施形態では、ホストLUは全てILUとして構成されるため、ホストLU構成テーブル20271のLU Typeは全て「ILU」となり、CLU Class、CLU Stripe Sizeは意味をなさない。

【0083】管理者は、管理端末5を操作してディスクアレイシステム構成管理手段70と通信し、ディスクアレイサブセット10のディスク容量、ディスクユニットの台数等の情報を得て、ディスクアレイサブセット10のLU110の設定、RAIDレベルの設定等を行う。次に管理者は、管理端末5によりディスクアレイシステム構成管理手段70と通信し、ディスクアレイスイッチ20を制御して、各ホスト30とディスクアレイサブセット20間の関係情報を設定する。

【0084】以上の操作により、ディスクアレイシステム1の構成が確立し、ホスト30から管理者が望む通りにLU110が見えるようになる。ディスクアレイ構成管理手段70は以上の設定情報を保存し、管理者からの操作に応じ構成の確認や、構成の変更を行うことができる。

【0085】本実施形態によれば、ひとたびディスクアレイシステム1を構成すれば、管理者からディスクアレイスイッチ20の存在を認識させることが無く、複数のディスクアレイサブシステムを1台のディスクアレイシステムと同様に扱うことができる。また、本実施形態によれば、ディスクアレイスイッチ20とディスクアレイサブセット10は、同一の操作環境によって統一的に操作することができ、その構成確認や、構成変更も容易になる。さらに、本実施形態によれば、従来使用していたディスクアレイシステムを本実施形態におけるディスクアレイシステムに置き換える場合に、ホスト30の設定を変更することなく、ディスクアレイシステム1の構成をそれまで使用していたディスクアレイシステムの構成に合わせることができ、互換性を維持できる。

【0086】[第4実施形態]以上説明した第1から第3の実施形態では、ホストI/Fにファイバチャネルを使用している。以下に説明する実施形態では、ファイバチャネル以外のインタフェースが混在した形態について説明する。

【0087】図17は、ホストI/FがパラレルSCSIである場合のホストI/Fノード203内部のIC2023の一構成例を示す。20230はパラレルSCSIのプロトコル制御を行うSCSIプロトコルコントローラ（SPC）、20233はファイバチャネルのプロトコル制御を行うファイバチャネルプロトコルコントローラ（FPC）、20231はパラレルSCSIとファイバチャネルのシリアルSCSIをプロトコル変換するプロトコル変換プロセッサ（PEP）、20232はプロトコル変換中データを一時保存するバッファ（BUF）である。

【0088】本実施形態において、ホスト30は、ディスクアレイI/Fノード203に対してSCSIコマンドを発行する。リードコマンドの場合、SPC20230は、これをBUF20232に格納し、PEP20231に割り込みでコマンドの受信を報告する。PEP20231は、BUF20232に格納されたコマンドを利用し、FPC20233へのコマンドに変換し、FPC20233に送る。FPC20233は、このコマンドを受信すると、フレーム形式に変換し、SC2022に引き渡す。この際、エクステンジID、シーケンスID、ソースID、デスティネーションIDは、以降の処理が可能ないようにPEP20231により付加される。あとのコマンド処理は、第1実施形態と同様に行われる。

【0089】ディスクアレイサブセット10は、データの準備が完了すると、データ転送準備完了フレームの発行、データ転送、正常終了後ステータスフレームの発行を実施する。ディスクアレイサブセット10からIC2023までの間では、フレームヘッダ401やフレームペイロード402が必要に応じ変換されながら、各種フレームの転送が行われる。IC2023のFPC20233は、データ転送準備完了フレームを受信し、続いてデータを受信してBUF20232に格納し、続けて正常に転送が終わったならば、ステータスフレームを受信し、PTP20231に割り込みをかけてデータの転送完了を報告する。PTP20231は、割り込みを受けると、SPC20230を起動し、ホスト30に対しデータ転送を開始するよう指示する。SPC20230はホスト30にデータを送信し、正常終了を確認するとPTP20231に対し割り込みで正常終了を報告する。

【0090】ここでは、ファイバチャネル以外のホストI/Fの例としてパラレルSCSIを示したが、他のインタフェース、例えば、メインフレームへのホストI/FであるESCON等に対しても同様に適用することが可能である。ディスクアレイスイッチ20のホストI/Fノード203として、例えば、ファイバチャネル、パラレ

ルSCSI、及びESCONに対応したホストI/Fノードを設けることで、1台のディスクレイシステム1に、メインフレームと、パーソナルコンピュータ、ワークステーション等のいわゆるオープンシステムの両方を混在させて接続することが可能である。本実施形態では、ディスクレイI/Fとしては、第1から第3実施形態と同様、ファイバチャネルを用いているが、ディスクレイI/Fに対しても任意のI/Fを使用することが可能である。

【0091】[第5実施形態]次に、ディスクレイシステム1の構成管理の方法について、第5実施形態として説明する。図18は、本実施形態のシステム構成図である。本実施形態では、ホスト30が4台設けられている。ホスト“#0”、“#1”とディスクレイシステム1の間のI/F30はファイバチャネル、ホスト“#2”とディスクレイシステム1の間は、パラレルSCSI (Ultra SCSI)、ホスト“#3”とディスクレイシステム1の間は、パラレルSCSI (Ultra2 SCSI) で接続されている。

【0092】パラレルSCSIのディスクレイスイッチ20への接続は第4実施形態と同様に行われる。ディスクレイシステム1は、4台のディスクレイサブセット30を有する。ディスクレイサブセット“#0”には4つの独立LU、ディスクレイサブセット“#1”には2つの独立LUがそれぞれ構成されている。ディスクレイサブセット“#2”と“#3”で1つの統合LUが構成されている。本実施形態では、第1実施形態と同様、ホスト30に対しディスクレイサブセット10を隠蔽し、ファイバチャネルのフレームを変換するものとする。各LUに割り当てられるLUNは、ディスクレイサブセット“#0”のLUから順に、LUN=0、1、2、・・・6までの7つである。

【0093】図19は、管理端末5の表示画面上に表示される画面の一例である。図は、ホストI/F31と各論理ユニット(LU)との対応を示した論理接続構成画面である。

【0094】論理接続構成画面50には、各ホストI/F31に関する情報3100、各LU110に関する情報11000、ディスクレイサブセット10とLU110の関係等が表示される。ホストI/F31に関する情報としては、I/F種類、I/F速度、ステータス等が含まれる。LU110に関する情報としては、格納サブセット番号、LUN、容量、RAIDレベル、ステータス、情報、等が表示される。管理者はこの画面を参照することで、容易にディスクレイシステム1の構成を管理することができる。

【0095】論理接続構成画面50上で、ホストI/FとLUの間に引かれている線は、各ホストI/F31を経由してアクセス可能なLU110を示している。ホストI/Fから線の引かれていないLU110に対して、

そのホストI/Fに接続するホスト30からはアクセスできない。ホスト30によって、扱うデータ形式が異なり、また使用者も異なることから、セキュリティ維持上、適切なアクセス制限を設けることが不可欠である。そこで、システムを設定する管理者が、この画面を用いて、各LU110とホストI/Fとの間のアクセス許可をあたえるか否かによって、アクセス制限を実施する。図において、例えば、LU“#0”は、ホストI/F“#0”および“#1”からアクセス可能であるが、ホストI/F“#2”、“#3”からはアクセスできない。LU“#4”は、ホストI/F“#2”からのみアクセス可能である。

【0096】このようなアクセス制限を実現するためアクセス制限情報は、ディスクレイシステム構成管理手段70からディスクレイスイッチ20に対して送信される。ディスクレイスイッチ20に送られたアクセス制限情報は、各ホストI/Fノード203に配信され、各ホストI/Fノード203のDCT2027に登録される。ホストにより、アクセスが制限されたLUに対するLU存在有無の検査コマンドが発行された場合、各ホストI/Fノード203は、DCT2027の検査を行い、検査コマンドに対し応答しないか、あるいは、エラーを返すことで、そのLUは、ホストからは認識されなくなる。LU存在有無の検査コマンドとしては、SCSIプロトコルの場合、Test Unit Readyコマンドや、Inquiryコマンドが一般に用いられる。この検査なしに、リード/ライトが実施されることはないため、容易にアクセスの制限をかけることが可能である。

【0097】本実施形態ではホストI/F31毎にアクセス制限をかけているが、これを拡張することで、ホスト30毎にアクセス制限をかけることも容易に実現できる。また、ホストI/F31、ホスト30、あるいは、アドレス空間を特定して、リードのみ可、ライトのみ可、リード/ライトとも可、リード/ライトとも不可といった、コマンドの種別に応じたアクセス制限をかけることもできる。この場合、アクセス制限情報としてホストI/F番号、ホストID、アドレス空間、制限コマンド等を指定してディスクレイスイッチ20に制限を設定する。

【0098】次に、新たなディスクレイサブセット10の追加について説明する。ディスクレイサブセット10を新規に追加する場合、管理者は、ディスクレイスイッチ20の空いているディスクレイI/Fノード202に追加するディスクレイサブセット10を接続する。つづけて、管理者は、管理端末5を操作し、論理接続構成画面50に表示されている「最新状態を反映」ボタン5001を押下する。この操作に回答して、未設定のディスクレイサブセットを表す絵が画面上に表示される(図示せず)。このディスクレイサブセットの絵が選択されると、ディスクレイサブセットの設

定画面が現れる。管理者は、表示された設定画面上で、新規に追加されたディスクアレイサブセットの各種設定を実施する。ここで設定される項目にはLUの構成、RAIDレベル等がある。続けて、図19の論理接続構成図の画面に切り替えると、新規ディスクアレイサブセットとLUが現れる。以降、ホストI/F31毎に対するアクセス制限を設定し、「設定実行」ボタン5002を押下すると、ディスクアレイスイッチ20に対し、アクセス制限情報、およびディスクアレイサブセット、LUの情報が転送され、設定が実行される。

【0099】各ディスクアレイサブセット10にLU110を追加する際の手順も上述した手順で行われる。また、ディスクアレイサブセット、およびLUの削除についてもほぼ同様の手順で行われる。異なる点は、管理者が各削除部位を画面上で選択して「削除」ボタン5003を押下し、適切な確認が行われたのち、実行される点である。以上のように、管理端末70を用いることで、管理者はディスクアレイシステム全体を一元的に管理できる。

【0100】〔第6実施形態〕次に、ディスクアレイスイッチ20によるミラーリングの処理について、第6実施形態として説明する。ここで説明するミラーリングとは、2台のディスクアレイサブセットの2つの独立LUにより二重書きをサポートする方法であり、ディスクアレイサブセットのコントローラまで含めた二重化である。従って、信頼性は、ディスクのみの二重化とは異なる。

【0101】本実施形態におけるシステムの構成は図1に示すものと同じである。図1に示す構成において、ディスクアレイサブセット“#0”と“#1”は全く同一のLU構成を備えており、この2つのディスクアレイサブセットがホスト30からは1つのディスクアレイとして見えるものとする。便宜上、ミラーリングされたディスクアレイサブセットのペアの番号を“#01”と呼ぶ。また、各ディスクアレイサブセットのLU“#0”とLU“#1”によってミラーリングペアが形成され、このLUのペアを便宜上、LU“#01”と呼ぶ。DCT2027のホストLU構成テーブル20271上でLU#01を管理するための情報は、CLU Classに「Mirrored」が設定され、LU Info.として、LU#0とLU#1に関する情報が設定される。その他の各部の構成は第1実施形態と同様である。

【0102】本実施形態における各部の動作は、第1実施例とほぼ同様である。以下、第1実施形態と相違する点について、ディスクアレイスイッチ20のホストI/Fノード203の動作を中心に説明する。図20は、本実施形態におけるライト動作時に転送されるフレームのシーケンスを示す模式図、図21、22は、ライト動作時におけるホストI/Fノード203による処理の流れを示すフローチャートである。

【0103】ライト動作時、ホスト30が発行したライトコマンドフレーム(FCP\_CMD)は、IC2023により受信される(図20の矢印(a):ステップ21001)。IC2023により受信されたライトコマンドフレームは、第1実施形態で説明したリード動作時におけるステップ20002 20005と同様に処理される(ステップ21002 - 21005)。

【0104】SC2022は、SP2021を使ってDCT2027を検索し、ミラー化されたディスクアレイサブセット“#01”のLU“#01”へのライトアクセス要求であることを認識する(ステップ21006)。SC2022は、FB2025上に、受信したコマンドフレームの複製を作成する(ステップ21007)。SC2022は、DCT2027に設定されている構成情報に基づいてコマンドフレームの変換を行い、LU“#0”とLU“#1”の両者への別々のコマンドフレームを作成する(ステップ21008)。ここで、LU“#0”を主LU、LU“#1”を従LUと呼び、コマンドフレームにもそれぞれ主コマンドフレーム、従コマンドフレームと呼ぶ。そして、両者別々にET2026にエクステンジ情報を格納し、ディスクアレイサブセット“#0”およびディスクアレイサブセット“#1”に対し作成したコマンドフレームを発行する(図20の矢印(b0)(b1):ステップ21009)。

【0105】各ディスクアレイサブセット“#0”、“#1”は、コマンドフレームを受信し、それぞれ独立にデータ転送準備完了フレーム(FCP\_XFER\_RDY)をディスクアレイスイッチ20に送信する(図20の矢印(c0)(c1))。ディスクアレイスイッチ20では、ホストI/Fノード203が、第1実施形態におけるリード動作のステップ20011 20013と同様の処理により転送されてきたデータ転送準備完了フレームを処理する(ステップ21011 - 21013)。

【0106】各ディスクアレイサブセットからのデータ転送準備完了フレームがそろった段階で(ステップ21014)、SC2022は、主データ転送準備完了フレームに対する変換を実施し(ステップ21015)、IC2023により変換後のフレームをホスト30に送信する(図20の矢印(d):ステップ21015)。

【0107】ホスト30は、データ転送準備完了フレームを受信した後、ライトデータ送信のため、データフレーム(FCP\_DATA)をディスクアレイスイッチ20に送信する(図20の矢印(e))。ホスト30からのデータフレームは、IC2023により受信されると(ステップ21031)、リードコマンドフレームやライトコマンドフレームと同様に、FB2025に格納され、CRC検査、フレームヘッダの解析が行われる(ステップ21032、21033)。フレームヘッダの解析結果に基づき、ET2026がSP2021により検索され、エクステンジ情報が獲得される(ステップ21034)。



【0108】SC2022は、ライトコマンドフレームのときと同様に複製を作成し(ステップ21035)、その一方をディスクアレイサブセット“#0”内のLU“#0”に、他方をディスクアレイサブセット“#1”内のLU“#1”に向けて送信する(図20の矢印(f0)(f1):ステップ21037)。

【0109】ディスクアレイサブセット“#0”、“#1”は、各々、データフレームを受信し、ディスクユニット104に対しそれぞれライトし、ステータスフレーム(FCP\_RSP)をディスクアレイスイッチ20に送信する。

【0110】SC2022は、ディスクアレイサブセット“#0”、“#1”それぞれからステータスフレームを受信すると、それらのステータスフレームから拡張ヘッダを外してフレームヘッダを再現し、ET2026からエクスチェンジ情報を獲得する(ステップ21041、21042)。

【0111】ディスクアレイサブセット“#0”、“#1”の両者からのステータスフレームが揃うと(ステップ21043)、ステータスが正常終了であることを確認のうえ、LU“#0”からの主ステータスフレームに対する変換を行い(ステップ21044)、従ステータスフレーム消去する(ステップ21045)。そして、IC2023は、正常終了を報告するためのコマンドフレームをホストに送信する(図20の矢印(h):ステップ21046)。最後にSP2021は、ET2026のエクスチェンジ情報を消去する(ステップ21047)。

【0112】以上でミラーリング構成におけるライト処理が終了する。ミラーリングされたLU“#01”に対するリード処理は、データの転送方向が異なるだけで、上述したライト処理とほぼ同様に行われるが、ライトとは異なり、2台のディスクアレイサブセットにリードコマンドを発行する必要はなく、どちらか一方に対してコマンドフレームを発行すればよい。たとえば、常に主LUに対してコマンドフレームを発行してもよいが、高速化のため、主/従双方のLUに対して、交互にコマンドフレームを発行するなどにより、負荷を分散すると有効である。

【0113】上述した処理では、ステップ21014、及びステップ21043で2台のディスクアレイサブセット“#0”、“#1”の応答を待ち、両者の同期をとって処理が進められる。このような制御では、双方のディスクアレイサブセットでの処理の成功が確認されてから処理が進むため、エラー発生時の対応が容易になる。その一方で、全体の処理速度が、どちらか遅いほうの応答に依存してしまうため、性能が低下するという欠点がある。

【0114】この問題を解決するため、ディスクアレイスイッチにおいて、ディスクアレイサブセットの応答を待たずに次の処理に進んだり、ディスクアレイサブセットのどちらか一方からの応答があった時点で次の処理に

進む「非同期型」の制御をすることも可能である。非同期型の制御を行った場合のフレームシーケンスの一例を、図20において破線矢印で示す。

【0115】破線矢印で示されるフレームシーケンスでは、ステップ21016で行われるホストへのデータ転送準備完了フレームの送信が、ステップ21009の処理の後、ディスクアレイサブセット10からのデータ転送準備完了フレームを待たずに実施される。この場合、ホストに送信されるデータ転送準備完了フレームは、ディスクアレイスイッチ20のSC2022により生成される(破線矢印(d'))。

【0116】ホスト30からは、破線矢印(e')で示されるタイミングでデータフレームがディスクアレイスイッチ20に転送される。ディスクアレイスイッチ20では、このデータフレームが一旦FB2025に格納される。SC2022は、ディスクアレイサブセット10からのデータ転送準備完了フレームの受信に応答して、データ転送準備完了フレームが送られてきたディスクアレイサブセット10に対し、FB2025に保持されたデータフレームを転送する(破線矢印(f0')(f1'))。

【0117】ディスクアレイスイッチ20からホスト30への終了報告は、双方のディスクアレイサブシステム10からの報告(破線矢印(g0')(g0'))があった時点でおこなわれる(破線矢印(h'))。このような処理により、図20に示される時間Taの分だけ処理時間を短縮することが可能である。

【0118】ディスクアレイスイッチ20とディスクアレイサブセット10間のフレーム転送の途中でエラーが発生した場合、以下の処理が実施される。

【0119】実行中の処理がライト処理の場合、エラーが発生したLUに対し、リトライ処理が行われる。リトライが成功すれば、処理はそのまま継続される。あらかじめ設定された規定の回数のリトライが失敗した場合、ディスクアレイスイッチ20は、このディスクアレイサブセット10(もしくはLU)に対するアクセスを禁止し、そのことを示す情報をDCT2027に登録する。また、ディスクアレイスイッチ20は、MP200、通信コントローラ204を経由して、ディスクシステム構成手段70にそのことを通知する。

【0120】ディスクシステム構成手段70は、この通知に回答して管理端末5にアラームを発行する。これにより管理者は、トラブルが発生したことを認識できる。その後、ディスクアレイスイッチ20は、正常なディスクアレイサブセットを用いて運転を継続する。ホスト30は、エラーが発生したことを認識することなく、処理を継続できる。

【0121】本実施形態によれば、2台のディスクアレイサブシステムでミラー構成を実現できるので、ディスクの耐障害性を上げることができ、また、ディス

クアレイドコントローラ、ディスクアレイド／Ｆ、及びディスクアレイド／Ｆノードの耐障害性を上げることができ、内部バスの二重化等することなくディスクアレイドシステム全体の信頼性を向上させることができる。

【0122】〔第7実施形態〕次に、3台以上のディスクアレイドサブセット10を統合し、1台の論理的なディスクアレイドサブセットのグループを構成する方法について説明する。本実施形態では、複数のディスクアレイドサブセット10にデータを分散して格納する。これにより、ディスクアレイドサブセットへのアクセスを分散させ、特定のディスクアレイドサブセットへのアクセスの集中を抑止することで、トータルスループットを向上させる。本実施形態では、ディスクアレイドスイッチによりこのようなストライピング処理を実施する。

【0123】図23は、本実施形態におけるディスクアレイドシステム1のアドレスマップである。ディスクアレイドサブセット10のアドレス空間は、ストライプサイズSでストライピングされている。ホストから見たディスクアレイドシステム1のアドレス空間は、ストライプサイズS毎に、ディスクアレイドサブセット“＃0”、“＃1”、“＃2”、“＃3”に分散されている。ストライプサイズSのサイズは任意であるが、あまり小さくない方がよい。ストライプサイズSが小さすぎると、アクセスすべきデータが複数のストライプに属するストライプまたぎが発生したときに、その処理にオーバーヘッドが発生するおそれがある。ストライプサイズSを大きくすると、ストライプまたぎが発生する確率が減少するので性能向上のためには好ましい。LUの数は任意に設定することができる。

【0124】以下、本実施形態におけるホストI／Fノード203の動作について、図24に示す動作フローチャートを参照しつつ第1実施形態との相違点に着目して説明する。なお、本実施形態では、DCT2027のホストLU構成テーブル20271上で、ストライピングされたホストLUに関する情報のCLU Classには「Striped」が、CLU Stripe Sizeにはストライプサイズ「S」が設定される。

【0125】ホスト30がコマンドフレームを発行すると、ディスクアレイドスイッチ20は、ホストI／Fノード203のIC2023でこれを受信する（ステップ2001）。SC2022は、IC2023からこのコマンドフレームを受け取り、SP2021を使ってDCT2027を検索し、ストライピングする必要があることを認識する（ステップ2005）。

【0126】次に、SC2022は、SP2021によりDCT2027を検索し、ストライプサイズSを含む構成情報から、アクセスの対象となるデータが属するストライプのストライプ番号を求め、このストライプがどのディスクアレイドサブセット10に格納されているか特定する（ステップ2006）。この際、ストライプまたぎ

が発生する可能性があるが、この場合の処理については後述する。ストライプまたぎが発生しない場合、SP2021の計算結果に基づき、SC2022はコマンドフレームに対し変換を施し（ステップ2007）、エクステンジ情報をET2026に格納する（ステップ2008）。以降は、第1実施形態と同様の処理が行われる。

【0127】ストライプまたぎが発生した場合、SP2021は、2つのコマンドフレームを生成する。この生成は、例えば、ホスト30が発行したコマンドフレームを複製することで行われる。生成するコマンドフレームのフレームヘッダ、フレームペイロード等は、新規に設定する。第6実施形態と同様、SC2022でコマンドフレームの複製を作成した後、変換を実施することも可能であるが、ここでは、SP2021により新規に作成されるものとする。SC2022は、2つのコマンドフレームが生成されると、これらを各ディスクアレイドサブセット10に送信する。

【0128】この後、第1実施形態と同様にデータ転送が実施される。ここで、本実施形態では、第1実施形態、あるいは第6実施形態と異なり、データ自体を1台のホスト30と2台のディスクアレイドサブセット10間で転送する必要がある。たとえば、リード処理の場合、2台のディスクアレイドサブセット10から転送されるデータフレームは、すべてホスト30に転送する必要がある。この際SC2022は、各ディスクアレイドサブセット10から転送されてくるデータフレームに対し、ET2026に登録されたエクステンジ情報に従い、適切な順番で、適切なエクステンジ情報を付加してホスト30に送信する。

【0129】ライト処理の場合は、コマンドフレームの場合と同様、2つのデータフレームに分割して、該当するディスクアレイドサブセット10に転送する。なお、データフレームの順序制御は、ホスト、あるいはディスクアレイドサブセットがアウトオブオーダー（Out of Order）機能と呼ばれる、順不同処理に対応しているならば必須ではない。

【0130】最後に、すべてのデータ転送が完了し、ディスクアレイドスイッチ20が2つのステータスフレームをディスクアレイドサブセット10から受信すると、SP2021（あるいはSC2022）は、ホスト30へのステータスフレームを作成し、これをIC2023によりホスト30に送信する。

【0131】本実施形態によれば、アクセスを複数のディスクアレイドサブセットに分散することができるので、トータルとしてスループットを向上させることができるとともに、アクセスレイテンシも平均的に低減させることが可能である。

【0132】〔第8実施形態〕次に、2台のディスクアレイドシステム（またはディスクアレイドサブセット）間における複製の作成について、第8実施形態として説明す

る。ここで説明するようなシステムは、2台のディスクアレイシステム的一方を遠隔地に配置し、天災等による他方のディスクアレイシステムの障害に対する耐性を備える。このような災害に対する対策をディザスタリカバリと呼び、遠隔地のディスクアレイシステムとの間で行われる複製の作成のことをリモートコピーと呼ぶ。

【0133】第6実施形態で説明したミラーリングでは、地理的にほぼ同一の場所に設置されたディスクアレイサブセット10でミラーを構成するので、ディスクアレイI/F21はファイバチャネルでよい。しかし、リモートコピーを行うディスクアレイ（ディスクアレイサブセット）が10kmを越える遠隔地に設置される場合、中継なしでファイバチャネルによりフレームを転送する事ができない。ディザスタリカバリに用いられる場合、お互いの間の距離は通常数百km以上となる、このため、ファイバチャネルでディスクアレイ間を接続することは実用上不可能であり、ATM (Asynchronous Transfer Mode) 等による高速公衆回線や衛星通信等が用いられる。

【0134】図25は、本実施形態におけるディザスタリカバリシステムの構成例である。

【0135】81はサイトA、82はサイトBであり、両サイトは、地理的な遠隔地に設置される。9は公衆回線であり、ATMパケットがここを通過する。サイトA81、およびサイトB82は、それぞれディスクアレイシステム1を有する。ここでは、サイトA81が通常使用される常用サイトであり、サイトB82はサイトA81が災害等でダウンしたときに使用されるリモートディザスタリカバリサイトである。

【0136】サイトA81のディスクアレイシステム10のディスクアレイサブセット“#0”、“#1”の内容は、サイトB82のディスクアレイシステム10のリモートコピー用ディスクアレイサブセット“#0”、“#1”にコピーされる。ディスクアレイスイッチ20のI/Fノードのうち、リモートサイトに接続するものはATMを用いて公衆回線9に接続されている。このノードをATMノード205と呼ぶ。ATMノード205は、図5に示すホストI/Fノードと同様に構成され、IC2023がATM-ファイバチャネルの変換を行う。この変換は、第4実施形態におけるSCSI-ファイバチャネルの変換と同様の方法により実現される。

【0137】本実施形態におけるリモートコピーの処理は、第6実施形態におけるミラーリングの処理と類似する。以下、第6実施形態におけるミラーリングの処理と異なる点について説明する。

【0138】ホスト30がライトコマンドフレームを発行すると、サイトA81のディスクアレイシステム10は、第6実施形態における場合と同様にフレームの二重化を実施し、その一方を自身のディスクアレイサブセット10に転送する。他方のフレームは、ATMノード20

5によりファイバチャネルフレームからATMパケットに変換され、公衆回線9を介してサイトB82に送られる。

【0139】サイトB82では、ディスクアレイスイッチ20のATMノード205がこのパケットを受信する。ATMノード205のIC2023は、ATMパケットからファイバチャネルフレームを再現し、SC2022に転送する。SC2022は、ホスト30からライトコマンドを受信したときと同様にフレーム変換を施し、リモートコピー用のディスクアレイサブセットに転送する。以降、データ転送準備完了フレーム、データフレーム、ステータスフレームのすべてにおいて、ATMノード205においてファイバチャネル-ATM変換を行い、同様のフレーム転送処理を実施することにより、リモートコピーが実現できる。

【0140】ホスト30がリードコマンドフレームを発行した際には、ディスクアレイスイッチ20は、自サイトのディスクアレイサブセット10に対してのみコマンドフレームを転送し、自サイトのディスクアレイサブセット10からのみデータをリードする。このときの動作は、第1実施形態と同一となる。

【0141】本実施形態によれば、ユーザデータをリアルタイムでバックアップし、天災等によるサイト障害、ディスクアレイシステム障害に対する耐性を備えることができる。

【0142】〔第9実施形態〕次に、一台のディスクアレイサブセット10に包含される複数のLUの統合について説明する。例えば、メインフレーム用のディスク装置は、過去のシステムとの互換性を維持するために、論理ボリュームのサイズの最大値が2GBに設定されている。このようなディスクアレイシステムをオープンシステムでも共用する場合、LUは論理ボリュームサイズの制限をそのまま受けることになり、小サイズのLUが多数ホストから見えることになる。このような方法では、大容量化が進展した場合に運用が困難になるという問題が生じる。そこで、ディスクアレイスイッチ20の機能により、この論理ボリューム（すなわちLU）を統合して一つの大きな統合LUを構成することを考える。本実施形態では、統合LUの作成をディスクアレイスイッチ20で実施する。

【0143】本実施形態におけるLUの統合は、第1実施形態における複数のディスクアレイサブセット10による統合LUの作成と同一である。相違点は、同一のディスクアレイサブセット10内の複数のLUによる統合であることだけである。ディスクアレイシステムとしての動作は、第1実施形態と全く同一となる。

【0144】このように、同一のディスクアレイサブセット10に包含される複数のLUを統合して一つの大きなLUを作成することで、ホストから多数のLUを管理する必要がなくなり、運用性に優れ、管理コストを低減

したディスクアレイシステムを構築できる。

【0145】[第10実施形態]次に、ディスクアレイスイッチ10による交代バスの設定方法について、図26を参照しつつ説明する。

【0146】図26に示された計算機システムにおける各部の構成は、第1の実施形態と同様である。ここでは、2台のホスト30が、各々異なるディスクアレイI/F21を用いてディスクアレイサブセット10をアクセスするとように構成していると仮定する。図では、ディスクアレイサブセット、ディスクアレイスイッチ20のホストI/Fノード203およびディスクアレイI/Fノード202は、ここでの説明に必要な数しか示されていない。

【0147】ディスクアレイサブセット10は、図2と同様の構成を有し、2つのディスクアレイI/Fコントローラはそれぞれ1台のディスクアレイスイッチ20に接続している。ディスクアレイスイッチ20の各ノードのDCT227には、ディスクアレイI/F21の交代バスが設定される。交代バスとは、ある一つのバスに障害が発生した場合にもアクセス可能になるように設けられる代替のバスのことである。ここでは、ディスクアレイI/F“#0”の交替バスをディスクアレイI/F“#1”、ディスクアレイI/F“#1”の交替バスをディスクアレイI/F“#0”として設定しておく。同様に、ディスクアレイサブセット10内の上位アダプタ間、キャッシュ・交代メモリ間、下位アダプタ間のそれぞれについても交代バスを設定しておく。

【0148】次に、図26に示すように、ディスクアレイサブセット1の上位アダプタ“#1”に接続するディスクアレイI/F21が断線し、障害が発生したと仮定して、交替バスの設定動作を説明する。このとき、障害が発生したディスクアレイI/F21を利用しているホスト“#1”は、ディスクアレイサブセット10にアクセスできなくなる。ディスクアレイスイッチ20は、ディスクアレイサブセット10との間のフレーム転送の異常を検出し、リトライ処理を実施しても回復しない場合、このバスに障害が発生したと認識する。

【0149】バスの障害が発生すると、SP2021は、DCT2027にディスクアレイI/F“#1”に障害が発生したことを登録し、交代バスとしてディスクアレイI/F“#0”を使用することを登録する。以降、ホストI/Fノード203のSC2022は、ホスト“#1”からのフレームをディスクアレイI/F“#0”に接続するディスクアレイI/Fノード202に転送するように動作する。

【0150】ディスクアレイサブセット10の上位アダプタ101は、ホスト“#1”からのコマンドを引き継いで処理する。また、ディスクアレイスイッチ20は、ディスクアレイシステム構成管理手段70に障害の発生を通知し、ディスクアレイシステム構成管理手段70に

より管理者に障害の発生が通報される。

【0151】本実施形態によれば、バスに障害が発生した際の交替バスへの切り替えを、ホスト側に認識させることなく行うことができ、ホスト側の交代処理設定を不要にできる。これにより、システムの可用性を向上させることができる。

【0152】以上説明した各実施形態では、記憶メディアとして、すべてディスク装置を用いたディスクアレイシステムについて説明した。しかし、本発明は、これに限定されるものではなく、記憶メディアとしてディスク装置に限らず、光ディスク装置、テープ装置、DVD装置、半導体記憶装置等を用いた場合にも同様に適用できる。

【0153】

【発明の効果】本発明によれば、計算機システムの規模、要求などに応じた記憶装置システムの拡張、信頼性の向上などを容易に実現することのできる記憶装置システムを実現することができる。

【図面の簡単な説明】

【図1】第1実施形態のコンピュータシステムの構成図である。

【図2】第1実施形態のディスクアレイサブセットの構成図である。

【図3】第1実施形態のディスクアレイスイッチの構成図である。

【図4】第1実施形態におけるディスクアレイスイッチのクロスバススイッチの構成図である。

【図5】第1実施形態におけるディスクアレイスイッチのホストI/Fノードの構成図である。

【図6】システム構成テーブルの構成図である。

【図7】サブセット構成テーブルの構成図である。

【図8】ファイバチャネルのフレームの構成図である。

【図9】ファイバチャネルのフレームヘッダの構成図である。

【図10】ファイバチャネルのフレームペイロードの構成図である。

【図11】ホストからのリード動作時にファイバチャネルを通して転送されるフレームのシーケンスを示す模式図である。

【図12】ホストLU、各ディスクアレイサブセットのLU、及び各ディスクユニットの対応関係を示す模式図である。

【図13】ライト処理時のホストI/Fノードにおける処理のフローチャートである。

【図14】スイッチングバケットの構成図である。

【図15】複数のディスクアレイスイッチをクラスタ接続したディスクアレイシステムの構成図である。

【図16】第2実施形態におけるコンピュータシステムの構成図である。

【図17】第4実施形態におけるディスクアレイスイッ

このインタフェースコントローラの構成図である。

【図18】第5実施形態におけるコンピュータシステムの構成図である。

【図19】論理接続構成画面の表示例を示す画面構成図である。

【図20】第6実施形態におけるフレームシーケンスを示す模式図である。

【図21】第6実施形態のミラーリングライト処理時の  
 ホストI/Fノードにおける処理のフローチャートである。

【図22】第6実施形態のミラーリングライト処理時の  
 ホストI/Fノードにおける処理のフローチャートである。

【図23】第7実施形態におけるホストLUと各ディスクアレイサブセットのLUとの対応関係を示す模式図で

ある。

【図24】第7実施形態におけるホストI/Fノードの処理を示すフローチャートである。

【図25】第8実施形態におけるディザスタリカバリシステムの構成図である。

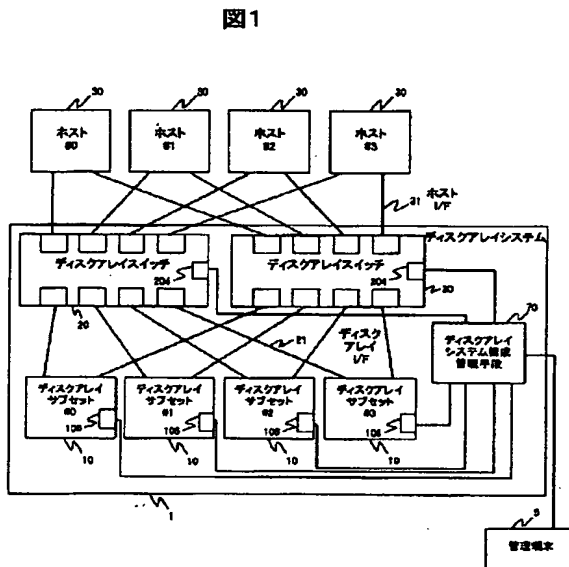
【図26】 交替パスの設定についての説明図である。

【符号の説明】

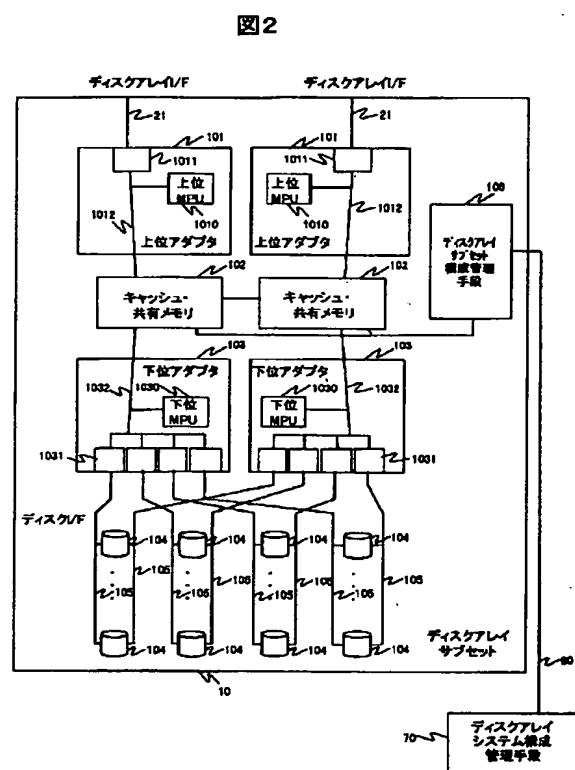
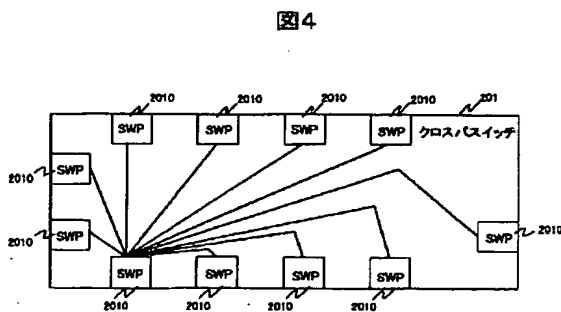
1…ディスクアレイシステム、5…管理端末、10…ディスクアレイサブセット、20…ディスクアレイスイッチ、30…ホストコンピュータ、70…ディスクアレイシステム構成管理手段、200…管理プロセッサ、201…クロスバススイッチ、202…ディスクアレイI/Fノード、203…ホストI/Fノード、204…通信コントローラ。

【図1】

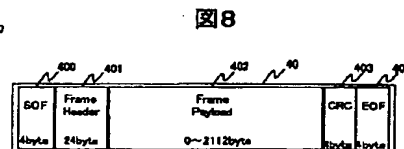
【図2】



【図4】

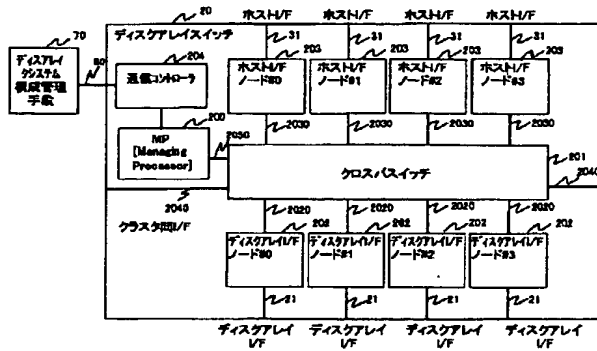


【図8】



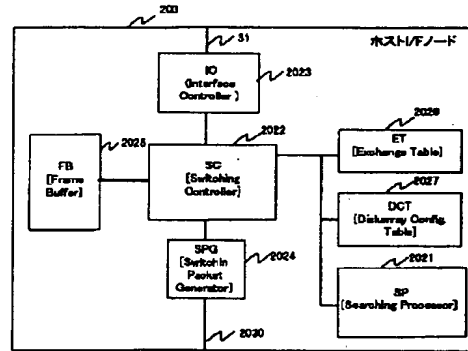
【図3】

図3



【図5】

図5



【図6】

図6

システム構成テーブル									
ホストLUテーブル									
Host-LU No.	LU Type	CLU Class	CLU Stripe Size	Condition	LU Info	LU Info	LU Info	LU Info	LU Info
0	CLU	Joined	—	Normal	#0 0 n0	#1 0 n1	#2 0 n2	#3 0 n3	—
1	—	—	—	Not Defined	—	—	—	—	—
2	—	—	—	Not Defined	—	—	—	—	—
...	...	...	...	...	...	...	...	...	...

ディスクアレイノード構成テーブル			
Subset	Subset Port No.	Switch No.	UF Node No.
#0	0	0	#0
#0	1	1	#0
#1	0	0	#1
#1	1	1	#1
...	...	...	...

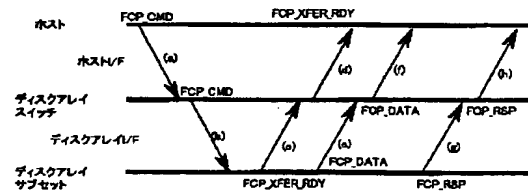
【図10】

図10

LUN (High)
LUN (Low)
CNTL
CDB (word0)
CDB (word1)
CDB (word2)
CDB (word3)
Data Length

【図11】

図11



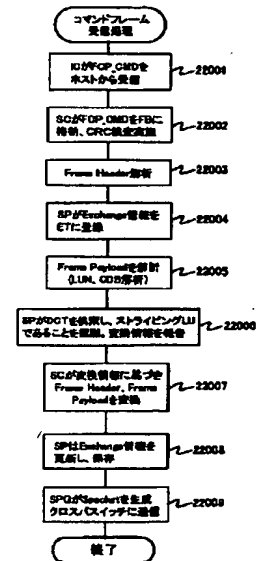
【図9】

図9

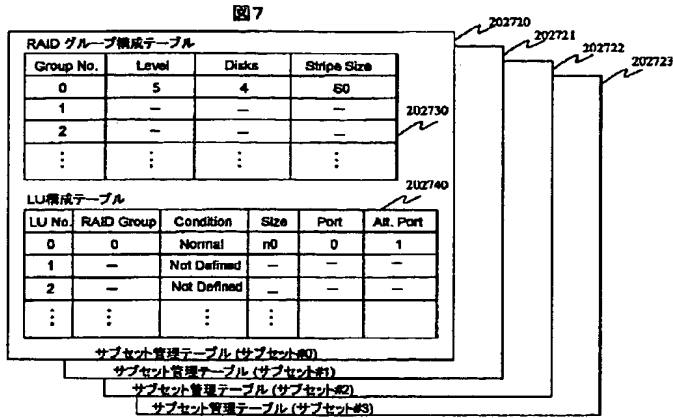
R_CTL	D_ID
未使用	S_ID
Type	F_CTL
SEQ_ID	DF_CTL
OX_ID	RX_ID
Parameters	

【図24】

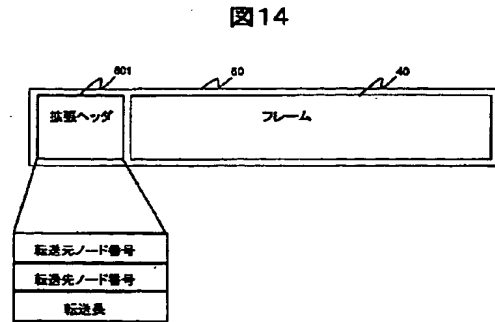
図24



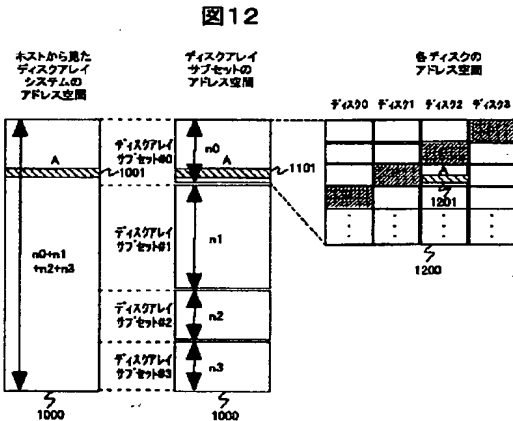
【図7】



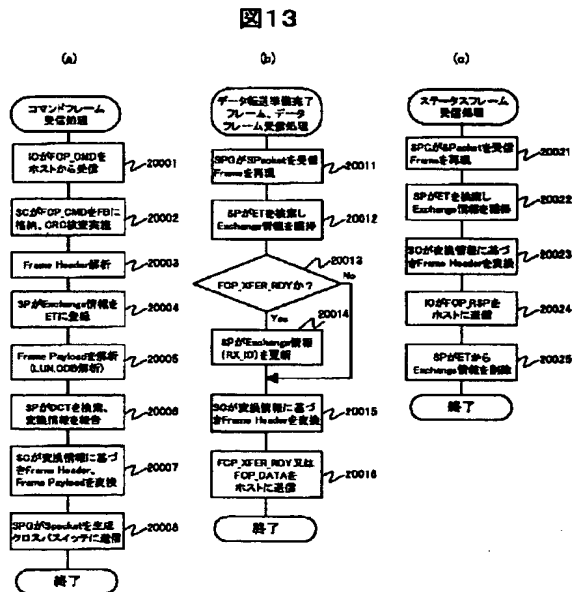
【図14】



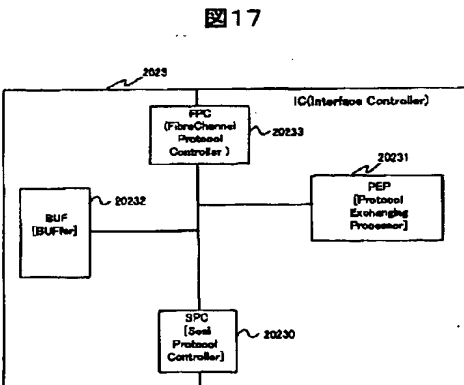
【図12】



【図13】

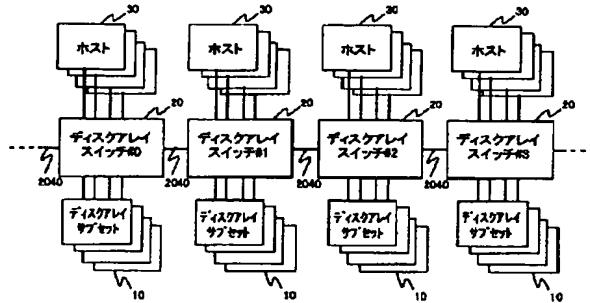


【図17】



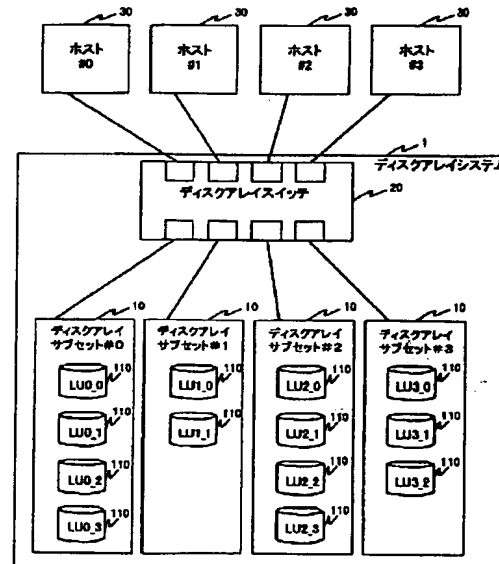
【図15】

図15



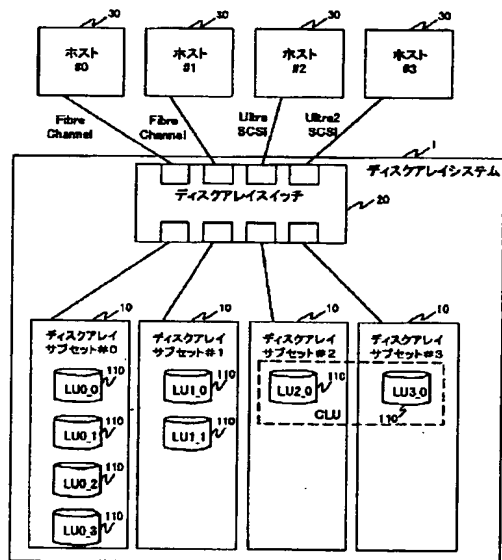
【図16】

図16



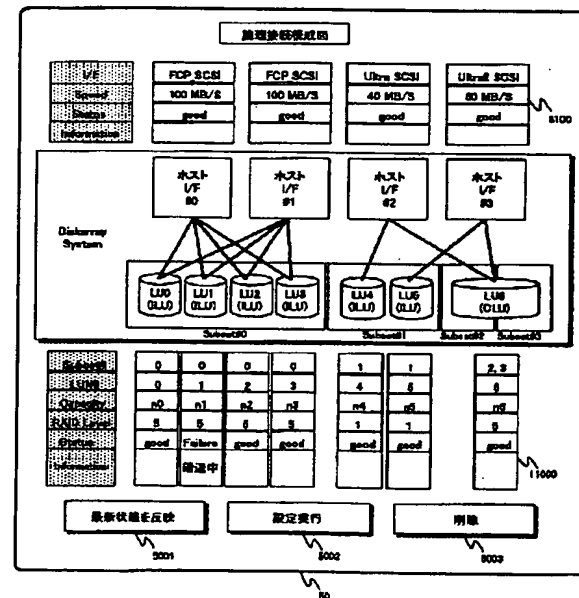
【図18】

図18



【図19】

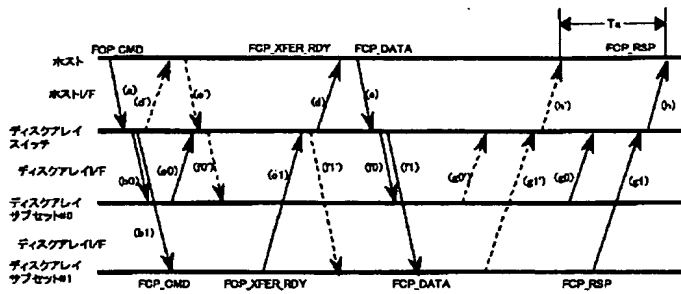
図19





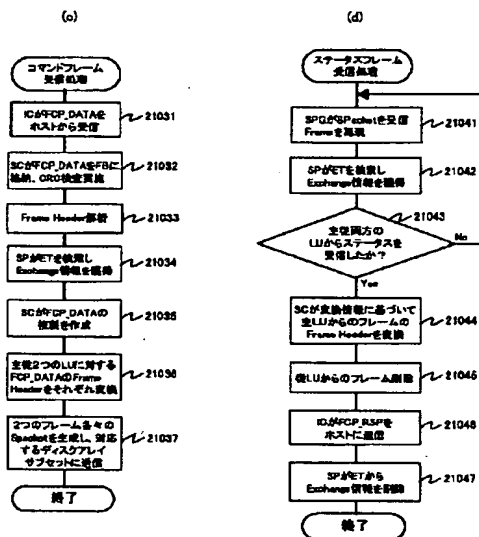
【図20】

圖20



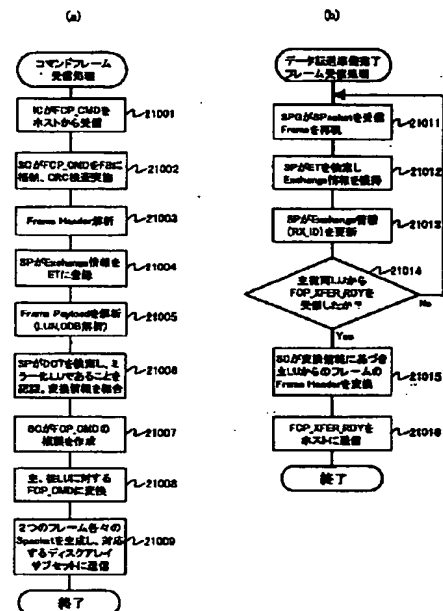
【图 22】

**圖22**



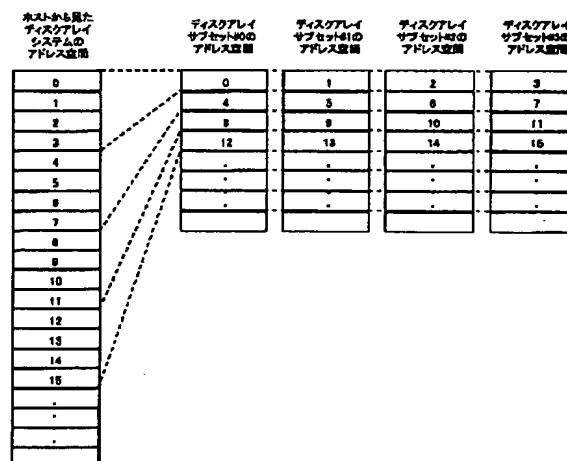
【図21】

**图21**



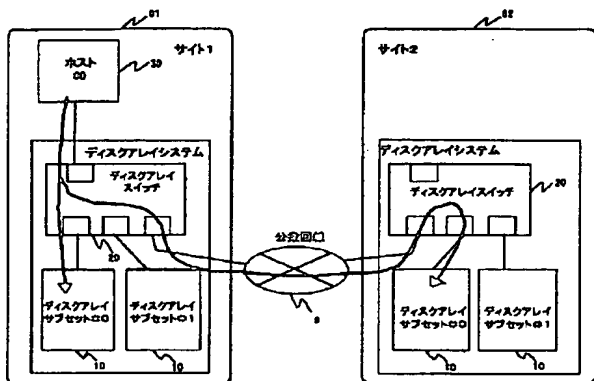
·【図23】

**圖 23**



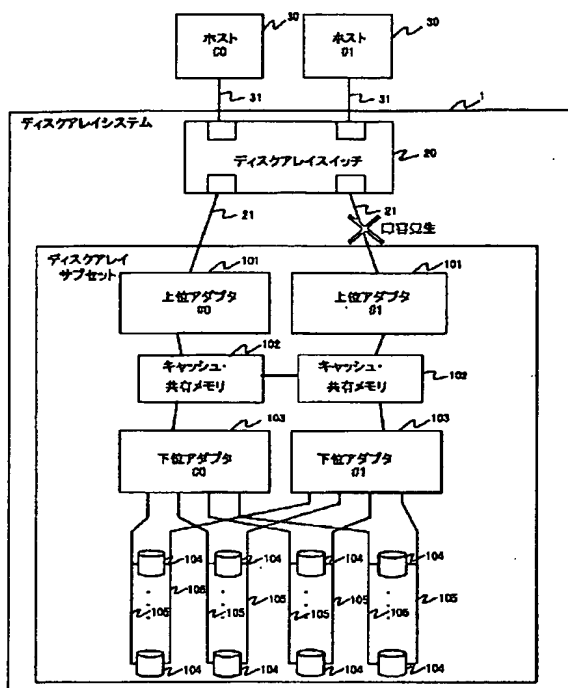
【图25】

**圖 25**



【図26】

**圖 26**



フロントページの続き

(72)発明者 山本 彰

神奈川県川崎市麻生区王禅寺1099番地 株  
 式会社日立製作所システム開発研究所内

(72) 発明者 味松 康行

神奈川県川崎市麻生区王禅寺1099番地 株式会社日立製作所システム開発研究所内

(72)発明者 佐藤 雅彦

神奈川県小田原市国府津2880番地 株式会社日立製作所ストレージシステム事業部内